Harnessing the Power of the Cloud: Revenue and Fairness Considerations in Multi-Resource Allocations

Carlee Joe-Wong and Soumya Sen

Electrical and Computer Engineering, Carnegie Mellon University

Information & Decision Sciences, Carlson School of Management, U of Minnesota

Emails: cjoewong@andrew.cmu.edu, ssen@umn.edu

Abstract

Cloud computing is a transformational technology that reduces upfront infrastructure costs, democratizes access to computing and storage capacity, and helps businesses innovate and compete in the digital economy. However, the distribution of these significant social and economic benefits among the various stakeholders in the cloud ecosystem will critically depend on pricing and resource allocation decisions made by cloud providers. These decisions impact not only the provider revenue, but also the fairness of resources allocated among cloud users and the overall social welfare. Several unique features of cloud computing (e.g., the presence of multiple, non-fungible resources and heterogenous users) make the interplay between these competing objectives hard to understand. To truly harness the power of cloud computing and to establish regulatory benchmarks, there is a need for frameworks that quantify how resource allocation and pricing choices affect different objectives. In this work, we introduce an analytical model that incorporates user utility, endogenous pricing, fairness among heterogeneous users, social welfare, and cloud provider revenue. We identify conditions under which

there is a tradeoff between fairness and revenue, and quantify the extent of this tradeoff by formulating an optimization problem in which a cloud provider maximizes its revenue, subject to constraints on the desired level of fairness. Our work provides an initial step towards developing regulatory policies to ensure an equitable distribution of the transformative benefits of the cloud.

Keywords: Cloud computing, fairness, revenue, resource allocation, cloud neutrality

1 Introduction

Cloud computing provides a transformative way for businesses to access servers, store data, deploy new services and scale up their resources on-demand. New network technologies like virtualization increasingly allow for better control and sharing of computational resources across multiple clients, resulting in a distinct trend towards the creation of large centralized datacenters that lies at the heart of cloud computing. Recent studies have found that businesses can realize an average of 22% savings due to cloud computing, with benefits of up to 48% in areas like innovation [11]; as many as 95% of digital businesses have taken advantage of these benefits and begun using public cloud computing facilities [25]. By reducing up-front provisioning costs and democratizing access to shared computing and storage facilities, cloud computing can significantly improve productivity in businesses and organizations, resulting in greater value for societies and economies.

Though cloud computing offers significant overall benefits, the *distribution* of these benefits among the various stakeholders in the cloud ecosystem depends crucially on various operational decisions of cloud providers, and in particular on how they price and allocate their available resources among their (likely heterogenous) clients or users. This resource allocation decision may incorporate several possible considerations, such as: how can providers maximize their revenue? How can they ensure "fairness" in the benefits that the different clients receive? Allocating resources with such objectives is also complicated by the fact that

clients can configure the resources of their virtual machines based on their jobs' requirements, and the provider's control over the resource allocation stems primarily from the prices that it charges per job[1] [2, 23]. This leads to another question: can a cloud provider use pricing to achieve fairness while maximizing its revenue?

Answering these questions in the cloud computing context is particularly non-trivial for four main reasons. First, the cloud is a *multi-resource* setting in which clients are *heterogenous* in the ratios of different resources that they require to derive utility from completed jobs (e.g., some jobs may be compute-intensive, while others may be memory-intensive). Secondly, the resources required per job (e.g., CPU cycles, memory, I/O bandwidth) are *non-fungible*, that is, a provider cannot flexibly compensate for a client's deficit of one resource by providing more of another. Thirdly, the resources that each job needs for completion must be consumed in *bundles*, that is, an a-la-carte combination of individual resources from different cloud vendors is typically infeasible (e.g., a computational job running on a provider's virtual machine will use both CPU cycles and memory from that provider's datacenter). Fourthly, the notion of *fairness* is quite nebulous in multi-resource contexts, and it is unclear how to incorporate it in resource allocation decisions. Clients' requests for bundles of non-substitutable resources mean that "fairness" must account for the amount of *each* resource received by each client, as well as whether the allocated resources match each client's requested bundle. These features are somewhat unique to the cloud in that they are rarely present simultaneously in flexible manufacturing settings. Additionally, a user is more likely to discover whether the resource provider is being fair or not with online performance measurement software than in traditional flexible manufacturing processes. In this work, we therefore aim to provide IS researchers and practitioners with an analytical framework for structured explorations of revenue and fairness in cloud computing contexts.

[1]For virtual machines, the commonly used hourly pay-per-machine price can be easily translated to a price per job by dividing it by the number of jobs that the VM can process per unit time, given its resource configuration.

This work primarily approaches the resource allocation problem from a cloud provider's perspective based on two real-world motivations. Firstly, evaluating the tradeoff between revenue and fairness is a natural consideration for many cloud providers. Although revenue maximization is usually thought of as providers' primary objective, there is a business impetus for a cloud provider to be fair in how it allocates its resources among its clients (users) so as to prevent them from experiencing resource starvation. Such starvation can potentially have long-term consequences for the reputation and loss in business of the cloud provider. In our surveys with members of the technical staff at three large US based cloud operators (enterprise, public, and government cloud operators), all of them identified fairness as an important metric to consider in cloud resource allocation[2].

Secondly, cloud providers are beginning to face calls for ensuring *cloud neutrality* [28,33], mirroring the net neutrality debate around Internet resource allocation [9,10,21]. Cloud neutrality refers to the practice of ensuring free, robust competition among cloud providers [14], which, like the net neutrality debate, requires addressing various facets of possible discriminatory and unfair practices. These include ensuring that providers do not favor any single user [9, 12], clients are not locked into a specific provider through use of proprietary APIs [13,43], and that discretionary resources are allocated fairly among tenants in the same SLA (service level agreement) class [28]. Just as no cloud provider should gain an unfair advantage by deploying proprietary APIs, no provider should be able to gain an advantage by favoring the jobs from some of its users while starving others. In our context, such neutrality can be realized in the form of regulatory policies that require a revenue-maximizing

[2]Our survey respondents reported that fairness across clients is always an important factor in their resource allocation decisions. One respondent stated that *"Enforcing fairness among multiple cloud instances is one of the most important problems in [our] internal cloud computing platform."* Another respondent further opined that *"Guaranteeing fairness is always the most important task. In the cloud industry, efficiency and revenue maximization is always the second-tier goal. That is the reason why most of the public clouds are having a very low hardware utilization rate."* Additionally, all of them agreed that internal policies enforcing fair resource allocations would likely help them avoid the same kind of regulatory scrutiny that broadband providers have faced in the net neutrality debate. These comments highlight that fairness is a desirable property even for profit-seeking cloud providers.

provider to also satisfy a constraint on the minimum level of utility for all clients, or some similar fairness criteria. We take this perspective in setting up an optimization framework for the cloud provider in this work.

Incorporating "fairness" considerations presents its own challenges, namely that it is not obvious which fairness criterion should be used. Thus, we first survey the fairness literature by introducing some basic concepts related to resource allocation from the domain of welfare economics. At a fundamental level, any resource allocation problem is one in which a central decision-maker (e.g., a cloud operations manager or a regulator) decides on an assignment of utilities to different users from the set of feasible utilities[3]. When this assignment is done so as to maximize the sum of the utilities across all the users, it is known as a *utilitarian* or *Benthamite*[4] allocation. Mathematically, the utilitarian criterion tries to find a feasible allocation that maximizes $\sum_i u_i$, where u_i is the utility derived by the ith user. This is the commonly used notion of *social welfare* ($W = \sum_i u_i$) in the economics literature. However, social welfare, as defined here from the perspective of utilitarianism, is often criticized for resulting in inequitable resource allocations [35, 39]: it is possible that a socially efficient outcome is achieved by giving most of the resources to a select few, while denying resources to others and causing them to suffer significantly. Additionally, a pricing or compensation mechanism may not exist to deal with the resulting inequity; e.g., when a cloud provider allocates any discretionary resources (i.e., leftover resources after tenant SLAs have been met) [28].

To address inequity in resource allocation, welfare economists have introduced a family of *social welfare functions*, given by $\sum_i \frac{{u_i}^{1-\alpha}}{1-\alpha}$, which is parameterized by a single scalar α that captures the decision-maker's aversion to inequity. This is also referred to as the *α-fairness scheme* [4]. Most well-known fairness measures can be generated from this family of functions

[3]The net utility realized by an individual user depend on the user's valuation, allocation, price, etc.

[4]Jeremy Bentham was an English philosopher who is regarded as the founder of modern utilitarianism.

by varying the value of α. For instance, when $\alpha = 0$, the social welfare function becomes the previously discussed utilitarian objective ($\sum_i u_i$), which lacks equity considerations [6, 18]. For $\alpha = 1$, the α-fairness scheme corresponds to *proportional fairness* [8], in which the objective function becomes $\sum_i log(u_i)$. Maximizing the proportional fairness is equivalent to maximizing the objective $\prod_i u_i$ in a *Nash bargaining* setting. Similarly, when $\alpha \to \infty$, the scheme converges to *Maximin fairness*, i.e., $min_i\ u_i$ is the measure of fairness [26, 34]. Maximin (max-min) fairness follows from *Rawls' theory*[5] of justice [35], which advocates the allocation of resources in a way that maximizes the utility of the least well-off. Thus, a higher value of α in the α-fairness scheme will lead to a fairer overall allocation [24, 29, 41]. These functions thus parameterize the tradeoff between fairness and social welfare: by increasing α from 0 to ∞, one can put more emphasis on fairness and less on social welfare.

Applying these fairness measures to a cloud setting is non-trivial. First, fairness can be measured on various metrics [24], e.g., user utilities realized, number of jobs completed, share of the dominant resource, etc. In this work, we follow the traditional approach in welfare economics, as described above, to consider fairness on user utilities[6], although our framework can be easily extended to consider alternative metrics. Second, the notions of fairness and equity are necessarily subjective, and hence no allocation scheme can objectively claim to be the "best" or the most fair. In this work, we focus on maximin fairness because it is often considered the "most fair" allocation [1, 7], as discussed above. This minimum utility requirement is also easily enforceable in practice by a regulator. We therefore adopt this fairness measure in our main model and explore the cloud provider's revenue optimiza-

[5]John Rawls was an American political philosopher who developed a theory of the Good as Justice and Justice conceived as Fairness.

[6]It bears mention that from an auditing standpoint, it is arguably easier to measure the realized user utilities as a function of the number of jobs completed for each user, than it is to precisely measure the internal allocation of resources among users in a highly virtualized datacenter. For example, Netflix and Amazon Prime both run on EC2. Netflix relies on the hypervisor underlying its Amazon EC2 instances to report on the "stolen time" (a measure of competition for CPU) to help identify incidents when its procured instances do not get enough CPU capacity. But if the hypervisor does not report this metric truthfully, it will be very hard to audit and account for such neutrality violations at the individual resource level.

tion problem under a maximin fairness constraint. Numerical evaluations are then used to demonstrate the robustness of the main results for other fairness criterion.

This paper makes four key contributions to the literature: first, we introduce a framework that endogenizes provider prices into resource allocation decisions, and can thus be used to reason about revenue and fairness in a cloud context. Second, we use this framework to quantify the exact combinations of user resource requirements under which a tradeoff between fairness and revenue exists. We find that the existence of the tradeoff depends on the degree of symmetry or asymmetry in users' resource requirements. Third, we provide an optimization formulation to quantify this fairness-revenue tradeoff for varying degrees of similarity in resource requirements of users. Lastly, this work provides initial insights into when regulatory intervention may be needed and how it can be implemented to ensure neutrality in the cloud.

This work is organized as follows: Section 2 provides an overview of the related research works and outlines our contributions to them. We then develop and analyze our model in Section 3 along with numerical evaluations of the model. We discuss our findings and conclude in Section 4. Proofs of all results are in the Appendix.

2 Literature Review

Our work draws upon several streams of research in welfare economics, computer science, and information systems, which we discuss next.

Welfare Economics: Much of the literature in welfare economics uses microeconomic theory to evaluate how a central decision maker should allocate resources among different users, given a constraint set of achievable utility allocations known as the utility possibility set [38]. A typical methodology to address the allocation problem involves the use of social welfare functions to rank feasible allocations of resources in terms of the social welfare they entail. Such functions typically include measures of both economic efficiency and equity. For

example, a widely used class of social welfare functions is the α-fairness scheme, given by $\sum_i \frac{{u_i}^{1-\alpha}}{1-\alpha}$, which is parameterized by a single parameter α that captures the decision-maker's aversion to inequity. Different values of α produce the most well-known realizations of social welfare functions, namely utilitarian ($\alpha = 0$), proportional ($\alpha \to 1$), and maximin fairness ($\alpha \to \infty$). We draw upon these definitions of fairness in our model, and our analysis uses the maximin fairness criterion [1,7]. The maximin criterion imposes a minimum utility guarantee for all users and can be easily enforced in practice by a cloud provider.

Cloud Computing: Much of the research on cloud computing in the computer science and information systems literature has focused on architecture, service models and their associated pricing challenges [46]. In particular, researchers have studied two main models for the pricing of cloud services: Infrastructure-as-a-service (IaaS) [40,45], in which raw resources such as CPU or memory storage are sold to end users; and Software-as-a-Service (SaaS) [30], which provides a more integrated platform that includes the use of software systems (e.g., data management applications). This study focuses solely on resource allocation in IaaS, as offered by third-party public cloud providers like Amazon's EC2 [2] and Google's Compute Engine [20].

Most works on IaaS pricing have focused on designing pricing strategies for a single resource scenario. For instance, [31] considers dynamic pricing of a single computing resource, subject to different customer budgets and service requirements (e.g., amount of computing time), while [50] considers a dynamic auction in which customers bid on resources in real time in order to complete their jobs, a format similar to Amazon's spot pricing [2]. Reviews of various pricing schemes are reported in [32, 37]. Other works (e.g. [15, 48]) have incorporated stochastic job arrivals and departures in their models of revenue-maximizing cloud providers. Wang et al. [47] consider pricing based on job completion times, and also account for electricity costs in the provider's optimization problem. All of these works largely deal with the operational aspects of cloud computing from the viewpoint of either a revenue-

maximizing cloud provider or a utility-maximizing bidding agent. Additionally, their focus is often on job scheduling and demand management as a means to improve the cloud's operational efficiency, as measured in terms of leftover capacity; thus, they take only the utilitarian perspective and have largely eschewed fairness considerations. In contrast, we account for the social welfare function to explore the impact of a cloud provider's operational decisions (e.g., multi-resource allocation) on the resulting fairness and revenue outcomes.

The notion of fairness has been primarily studied in the Internet and cloud computing literature in the context of single resource allocation [29]. For example, [5, 27, 41] consider the problem of "efficient" allocation of available link bandwidth to network flows so as to maximize throughput without attention to equitability of each individual flow. Even multi-resource allocation problems, such as scheduling jobs in a datacenter, have been treated as a single resource problem (e.g. the Hadoop and Dryad schedulers [49]). Only recently has the computer science community started to systematically examine the unique challenges of defining and enforcing fairness in the cloud's multi-resource setting. For instance, [16, 17] generalized the max-min fairness measure to multiple resource settings by introducing the notion of dominant resource fairness, while [24] introduced the notion of fairness on jobs. However, these works only model the *supply-side* of providers' allocation decisions and focus on quantifying the fairness of a realized allocation under the capacity constraints for different types of resources. They do not model user utility and do not endogenize the cloud provider's pricing decision. In contrast, our framework incorporates a demand-side model and also incorporates the traditional economic perspective of computing fairness on the user utilities.

Lastly, this work also contributes to the emerging literature on "cloud neutrality" [19, 28, 36]. Kesidis et al. [28] advocate that *"with the public cloud providers poised to become indispensable utility providers, neutrality-related mandates will likely emerge to ensure a level playing field among their customers (tenants)."* These works have highlighted the need for enforcing fairness between affiliates of the cloud provider and its tenants; e.g., in how Ama-

zon's AWS resources are allocated between services like Netflix and Amazon Prime which both run on Amazon EC2. Similarly, cloud neutrality would require guarantees on how the cloud allocates any discretionary resources among tenants with the same SLA class. The framework we present in this work helps identify the conditions under which a tradeoff between fairness and revenue exists (as a function of the (a)symmetry in the ratios of resources required by the different clients), and hence, can inform future policy decisions on when and how a regulator may need to intervene to ensure fairness in the cloud. This work thus complements and extends the recent research in the IS community on net neutrality issues [9, 21] to the cloud computing context[7].

Tradeoff Analysis: As discussed earlier, the utilitarian objective maximizes efficiency but is neutral to fairness. A few recent works [6] have therefore undertaken worst-case analyses by comparing the value of the utilitarian objective achieved under some fair allocation to the value achieved by an allocation that maximizes the utilitarian objective (i.e., price of fairness). In a follow up work, Bertsimas et al. [7] quantify the tradeoff between efficiency and fairness when different users have different utility functions. Joe-Wong et al. [24] study a similar tradeoff between fairness and efficiency, where the efficiency metric is quantified in terms of the leftover (unused) resource capacity after a fair allocation. Our model differs from these studies in three key ways. First, we incorporate a demand-side formulation and endogenize the provider's pricing decision, which affects the achieved fairness. Second, we show that even if the users have the same utility function, but differ in the ratios of the different resources they need to derive utilities from completed jobs, a tradeoff between fairness and revenue can arise, even with endogenous pricing. Third, we derive conditions under

[7]While the focus of the net neutrality literature has been on a single-resource allocation (e.g., bandwidth), packet prioritization, tiering, and openness of service provider platforms that serve a two-sided market, the initial focus of cloud neutrality has primarily been on multi-resource allocation, with little consideration of prioritization or tiering. Additionally, a typical cloud provider is not a two-sided platform. But many of the issues raised in the net neutrality debate, such as openness and fair competition, may become relevant even in the cloud context.

which a tradeoff between fairness and revenue exists and solve for the cloud provider's profit maximization objective under a fairness constraint. Thus, this work takes an initial step towards developing a holistic model to evaluate the outcomes in the cloud's multi-resource allocation setting. In doing so, we also address the call by Tilson et al. [42] to put the study of digital infrastructures at the core of the IS research agenda.

3 Analysis

Proofs of all propositions and corollaries may be found in the Appendix. Proofs of the lemmas are given in our online technical report [3].

3.1 Cloud Computing: A Multi-resource Allocation Problem

Before introducing a formal analytical model, we first present a motivating toy example to evaluate the most basic question: Is there a tradeoff between revenue maximization and fairness in multi-resource allocation?

Consider an example in which two users have jobs to run in the cloud. Each user $i = 1, 2$ has a standard concave utility function, say $U_i(x_i) = (bx_i - x_i^2) - p_i x_i$, which captures user i's utility from running x_i number of jobs[8], where b is a fixed constant and p_i is the price per job [9] for user i. Then the user's demand function that optimizes the user utility will be $x_i^\star(p_i) = (b - p_i)/2$. We can thus find the revenue $p_1 x_1^\star + p_2 x_2^\star$ and maximin fairness[10] $\min\{U_1(x_1^\star(p_1)), U_2(x_2^\star(p_2))\}$ achieved by the cloud provider for a given set of prices.

The cloud provider choses these prices so as to satisfy the constraints on the capacity of

[8] We assume that users can derive utility from fractional numbers of jobs, e.g., if the jobs are divided into many individual tasks, as in MapReduce scenarios often considered in the computer science literature [16].

[9] These jobs are processed in virtual machines (VM) that are configured with ratios of CPU and memory capacity as needed by the type of job. The users pays at a unit rate per VM (e.g., in Google's Compute Engine), which can be directly translated to a per job price. It also bears mention that unlike resource pricing, differential pricing per job need not depend on the user's per-job resource requirements in any systematic way, and therefore represents the maximum flexibility for a provider to use pricing as a lever to overcome inequity.

[10] As alluded to earlier, in this work we primarily consider maximin fairness as the benchmark in our analysis as it is considered to be the most fair allocation [7].

resources available in the cloud. For numerical evaluation of the above setup, let us assume some real values for the parameters. First consider the case in which user 1 needs equal amounts of two resources, CPU cycles and memory, to complete each job, e.g., (0.5, 0.5) units, and user 2 needs (0.9, 0.1) units of CPU and memory respectively to complete each (computationally-intensive) job. Assume that all users configure their VMs to consume the exact bundle of resources specified by their per-job resource requirements. Our CPU and memory resource constraints are then $0.5(b-p_1)/2 + 0.9(b-p_2)/2 \leq 1$, $0.5(b-p_1)/2 + 0.1(b-p_2)/2 \leq 1$. For $b = 4$, the revenue-maximizing prices are $(p_1, p_2) = (2.3774, 2.6792)$, yielding a revenue of $p_1 x_1^\star + p_2 x_2^\star = 3.7$ and fairness of $\min\{U_1(x_1^\star(p_1)), U_2(x_2^\star(p_2))\} = 0.4361$. Now consider if the provider had instead chosen a different set of prices, say $(p_1, p_2) = (2.5714, 2.5714)$; in this case, the resulting revenue would drop only slightly to 3.67, but the fairness would increase to 0.5102, a 17% increase. Thus, the example demonstrates that there is a tradeoff between the achieved fairness and revenue that varies with the decisions taken by the cloud operator, and navigating this tradeoff carefully can help achieve greater fairness without significantly sacrificing revenue.

In the remainder of this section, we examine this fairness-revenue tradeoff in a general analytical framework. We first investigate the resource requirements under which a tradeoff arises (Section 3.2) and then present a principled method for a cloud provider to navigate this tradeoff (Section 3.3).

3.2 On Existence of the Fairness-Revenue Tradeoff

To find the exact conditions under which a tradeoff between fairness and revenue exists, we model two users[11] who wish to process jobs at a cloud provider. We develop a demand-side model in which users' resource requirements for jobs are denoted by (r_{11}, r_{12}) and (r_{21}, r_{22}), where r_{ij} is user i's per-job requirement for resource j and $r_{11}, r_{12}, r_{21}, r_{22} > 0$ (i.e., each

[11]The model can also be extended to consider two types of users who have different resource requirements (i.e., two types of jobs). For example, one type of users can be those with computationally-intensive jobs and the other with memory-intensive jobs.

user needs some amount of each resource to complete a job). These resources are non-substitutable (e.g., CPU cycles and IO bandwidth) and users are heterogenous in the ratios of the resources needed to complete their jobs (e.g., one type of job is CPU-intensive and another type can be IO-intensive). For example, MapReduce jobs that reads large volumes of data from disk or from the network (e.g., sorting, indexing, grouping, data importing and exporting, data transformation) are IO-intensive. On the other hand, jobs that processes data (e.g., clustering, complex text mining, natural-language processing, feature extraction, video encoding) are CPU-intensive. The price per job offered to the user i by the cloud provider is denoted by p_i. The utility (V_i) that user i receives depends on the number of jobs processed (x_i), given the price per job (p_i). We denote this utility with a generic isoelastic utility function[12], $V_i = \frac{cx_i^{1-\gamma}}{1-\gamma}$, where the parameter γ parameterizes the concavity of the utility function (i.e., quantifying the degree of diminishing marginal utility) and $c > 0$ is a positive constant that scales the utility level relative to the cost. We assume that users are homogeneous in the shape (i.e., have the same γ) of their utility functions[13]. Users choose the number of jobs that they process so as to maximize their surplus or net user utility, which is given by:

$$U_i = V_i - p_i x_i = \frac{cx_i^{1-\gamma}}{1-\gamma} - p_i x_i, \tag{1}$$

Users can be charged different prices because of their needs for different resource combinations to complete a job. We assume that users configure their virtual machines to consume bundles of resources that matches their jobs' resource requirements; users thus consume, and

[12]Isoelastic utility functions represent a large class of utility functions, parameterized by γ, that are most commonly used in the economics literature (e.g., $\gamma = 0$ corresponds to a linear utility model and $\gamma \to 1$ leads to a logarithmic utility model). Moreover, because such a constant elasticity function is concave and increasing, it captures diminishing marginal utility. It bears mention that the individual-level isoelastic utility functions are not to be confused with the $\alpha-$fairness social welfare function.

[13]This is arguably a more conservative setting for demonstrating any tradeoff between revenue and fairness perspectives. This is because any heterogeneity in the user utility functional forms is more likely to create these tradeoffs, but we show that such a tradeoff can exist even with homogeneous users who have the same utility functional form (but only differ in the ratios of different multiple resources that they need to complete their jobs).

are charged based on, the numbers of jobs. They choose the number of jobs to process, x_i, so as to maximize their utility functions (1), yielding the demand functions

$$x_i = x^\star(p_i) = c^{\frac{1}{\gamma}} p_i^{\frac{-1}{\gamma}}, \quad U^\star(r_i) = \frac{\gamma}{1-\gamma} c^{\frac{1}{\gamma}} p_i^{1-\frac{1}{\gamma}}. \tag{2}$$

Without loss of generality, we normalize the units of each resource so that the provider has a capacity of 1, yielding the resource constraints

$$r_{11}x^\star(p_1) + r_{21}x^\star(p_2) \leq 1, \quad r_{12}x^\star(p_1) + r_{22}x^\star(p_2) \leq 1. \tag{3}$$

We now consider how the prices p_1, p_2 are chosen. As discussed above, a cloud provider and a regulator would likely have different objectives in choosing the prices: the provider would likely choose prices to maximize its revenue,[14] while a regulator would be more concerned with fairness. We denote the revenue and (maximin) fairness[15] measures as

$$R(p_1, p_2) = p_1 x^\star(p_1) + p_2 x^\star(p_2), F(p_1, p_2) = \min\{U^\star(p_1), U^\star(p_2)\} \tag{4}$$

respectively, where $x^\star$ and $U^\star$ are given by (2).

As an alternative to maximin fairness, we can also use the utilitarian criteria, i.e., measuring the total benefits that the cloud provider and users jointly realize from a given set of

[14]Cloud providers could instead maximize their profits by including a cost term in their objectives. We suppose that the majority of the provider's cost lies in the fixed cost of provisioning cloud infrastructure capacity instead of processing jobs, allowing us to abstract away from operational costs like electricity, scheduling, and partial server shutdowns to instead constrain the amount of each available resource.

[15]For the remainder of the paper, we will simply use the term 'fairness' to refer to maximin fairness unless otherwise noted.

resource prices. In this case, the social welfare[16] is given by

$$W(p_1, p_2) = U_1^\star(p_1) + U_2^\star(p_2) + R(p_1, p_2) = \frac{c^{\frac{1}{\gamma}}}{1-\gamma}\left(p_1^{1-\frac{1}{\gamma}} + p_2^{1-\frac{1}{\gamma}}\right). \tag{5}$$

Substituting the optimal demands $x^\star$ from (2) into the expressions for W and R, we see that social welfare is a scalar multiple of cloud provider revenue, leading to the following lemma:

Lemma 1. *For general isoelastic user utility functions defined in (1), $W(p_1, p_2) = R(p_1, p_2)/(1-\alpha)$. Thus, the prices $(p_1^\star, p_2^\star)$ at which the cloud provider can maximize its revenue R subject to the resource constraints (3) also maximize the (utilitarian) social welfare W subject to the same constraints.*

Lemma 1 implies that there is no tradeoff between revenue and social welfare in this setting[17]. However, we may have a tradeoff between *fairness* and revenue; thus, the provider's and the regulator's objectives may not be aligned if the regulator uses fairness as a performance benchmark. We examine the circumstances of this fairness-revenue tradeoff in the discussion below.

We first identify under conditions for which is there no tradeoff between maximizing revenue and maximin fairness, i.e., there exists a set of prices (p_1, p_2) that maximizes the provider's revenue and the minimum utility across users. Without loss of generality, we assume the two resources are indexed (labeled) so that $r_{11} + r_{21} \geq r_{12} + r_{22}$ throughout this section. We then find conditions under which fairness is maximized:

Lemma 2. *The prices $p_1 = p_2 = c(r_{11} + r_{21})^\alpha$ are the unique maximizers of the fairness function $F = \min\{U^\star(p_1), U^\star(p_2)\}$ in (4), subject to the resource constraints (3).*

[16]This is same as the notion of total welfare, realized as the sum of user and provider utilities, from a utilitarian fairness perspective.

[17]In Section 3.4 we provide additional numerical evaluation of this tradeoff with non-isoelastic user utility functions.

Thus, at the fair solution, users are charged equal prices and can process equal numbers of jobs, resulting in equal utilities. Next we find the necessary conditions for the cloud provider to maximize its revenue.

Lemma 3. *Suppose that p_1, p_2 maximize the cloud provider's revenue. Then*

$$p_1 = \frac{\mu r_{11} + \nu r_{12}}{2}, \quad p_2 = \frac{\mu r_{21} + \nu r_{22}}{2} \tag{6}$$

for some $\mu, \nu \geq 0$ such that

$$\begin{aligned} \mu \left(r_{11} x^\star \left(\frac{\mu r_{11} + \nu r_{12}}{2} \right) + r_{21} x^\star \left(\frac{\mu r_{21} + \nu r_{22}}{2} \right) \right) &= \mu \\ \nu \left(r_{12} x^\star \left(\frac{\mu r_{11} + \nu r_{12}}{2} \right) + r_{22} x^\star \left(\frac{\mu r_{21} + \nu r_{22}}{2} \right) \right) &= \nu. \end{aligned} \tag{7}$$

We can now quantify when the fairness-revenue tradeoff arises by identifying conditions under which the fairness-maximizing prices of Lemma 2 do not satisfy the revenue-maximizing conditions of Lemma 3:

Proposition 1. *There is a fairness-revenue tradeoff if $r_{11} r_{22} \neq r_{12} r_{21}$, $r_{11} \neq r_{21}$, and $r_{11} + r_{21} \neq r_{12} + r_{22}$, i.e., the revenue-maximizing prices do not maximize fairness.*

In general, for most combinations (ratios) of resource requirements, these conditions are likely to be satisfied; hence, there is usually a fairness-revenue tradeoff that needs to be considered. However, if the resource requirements of the users are sufficiently aligned such that any one of the conditions in Proposition 1 is not met, then this tradeoff will not arise. We therefore need to formally characterize this "alignment" of resource requirements for which a tradeoff does not exist. Proposition 1 shows that there are at least three different cases of this alignment to consider, namely ($r_{11} r_{22} = r_{12} r_{21}$, $r_{11} = r_{21}$, and $r_{11} + r_{21} = r_{12} + r_{22}$).

Proposition 2. *Suppose without loss of generality that $r_{11} \geq r_{12}$ (we can re-index the users to ensure that this constraint is satisfied). There is no fairness-revenue tradeoff if and only if $r_{11} = r_{21}$ or $r_{11} + r_{21} = r_{12} + r_{22}$ and either $r_{11} \geq r_{22} \geq r_{12}$ or $r_{22} \geq r_{11} \geq r_{12}$.*

These conditions are usually satisfied in only a few specific cases, as $r_{11} = r_{21}$ and $r_{11}+r_{21} = r_{12}+r_{22}$ each eliminate one degree of freedom in choosing the four resource requirements. We can broadly categorize these two conditions as ensuring sufficient "symmetry" or "asymmetry" in the users' resource requirements respectively, either of which can ensure that charging users equal prices (i.e., maximizing fairness) also maximizes provider revenue, thus eliminating the tradeoff. Formally, we define "perfectly symmetric" resource requirements as the case when both users have identical resource needs[18] (i.e., $r_{11} = r_{21} \geq r_{12} = r_{22}$), and "perfectly asymmetric" resource requirements as the case when user 1's requirement for resource 1 equals user 2's requirement for resource 2, and user 1's requirement for resource 2 equals user 2's requirements for resource 1 (i.e., $r_{11} = r_{22} > r_{12} = r_{21}$ or $r_{11} = r_{22} < r_{12} = r_{21}$).

Proposition 2 shows that perfect symmetry and perfect asymmetry are sufficient, but not necessary, conditions for there to be no tradeoff between fairness and revenue. We can easily check that the conditions for perfect symmetry satisfy Proposition 2. However, if $r_{11} = r_{21} \geq r_{22} > r_{12}$, i.e., both users have the same requirements for resource 1, which exceed their (unequal) requirements for resource 2, then there would still be no tradeoff between fairness and revenue. In this case, the users' resource requirements are not perfectly symmetric, but are symmetric "enough" for there to be no fairness-revenue tradeoff.

Similarly, users with sufficiently asymmetric resource requirements would not experience any tradeoff between fairness and revenue. Perfectly asymmetric users, whose requirements

[18]Thus, if the users are homogeneous, the fairness-revenue tradeoff is avoidable by choosing the right prices. This shows that the cloud context, which features heterogenous user requirements in the ratios of the resources, is itself the driver for this tradeoff. That is why, such a tradeoff is not usually noticed in single-resource pricing and allocation settings (for users with same utility functional form).

can be written as $r_{11} = r_{22} = r$, $r_{12} = r_{21} = \beta r$ for some scalar $\beta \geq 0$, satisfy the conditions $r_{11} + r_{21} = r_{12} + r_{22}$ and $r_{11} \geq r_{22} \geq r_{12}$ in Proposition 2; thus, they do not have a tradeoff. However, Proposition 2 requires only the weaker condition that $r_{11} + r_{21} = r_{12} + r_{22}$ and $r_{11} \geq r_{22} \geq r_{12}$ or $r_{22} \geq r_{11} \geq r_{12}$ in order for there to be no tradeoff. These conditions ensure that user 1's requirement for resource 1 dominates its requirement for resource 2, and vice versa for user 2, but are weaker than perfect asymmetry. For instance, we might assume $r_{11} = 0.5$, $r_{22} = 0.4$; then $r_{12} = 0.1$ and $r_{21} = 0.2$ ensures that the resource requirements are sufficiently, though not perfectly, asymmetric to ensure that there is no fairness-revenue tradeoff.

The above analysis shows that a fairness-revenue tradeoff usually does not arise when the ratios of the resource requirements of users are sufficiently symmetric or sufficiently asymmetric (see exact conditions in Proposition 2). In such cases, there is little reason for regulatory intervention because the cloud provider's profit maximization also maximizes the (maximin) fairness criteria. It is in scenarios with other combination of resource requirements that may need a regulator to step in to require minimum fairness guarantees for all users.

3.3 Quantifying the Fairness-Revenue Tradeoff

Given the scenarios under which a fairness-revenue tradeoff exists, we now examine the extent of this tradeoff. Such results are of particular interest to cloud regulators who might wish to ensure a certain level of fairness without unduly impacting the cloud provider's revenue. For instance, the regulator may mandate that the achieved fairness exceed a given threshold ϕ; cloud providers would then price their resources so as to maximize their revenue while satisfying this constraint. We can thus model the outcome of the cloud provider's resource

allocation as a solution to the optimization problem

$$\begin{aligned} &\max_{p_1,p_2} p_1 x^\star(p_1) + p_2 x^\star(p_2) \\ &\quad \text{s.t.} \min\left\{U_1^\star\left(p_1\right), U_2^\star\left(p_2\right)\right\} \geq \phi \\ &\qquad r_{11}x_1^\star(p_1) + r_{21}x_2^\star(p_2) \leq 1,\ r_{12}x_1^\star(p_1) + r_{22}x_2^\star(p_2) \leq 1. \end{aligned} \tag{8}$$

The last two constraints represent the resource constraints, while the first constraint is the fairness constraint imposed by the regulator. The regulator then faces a question of how to choose the threshold ϕ. While a larger threshold will enforce a more fair allocation, choosing ϕ to be too large may significantly lower the cloud provider's revenue; the regulator would likely wish to achieve a balance between these two quantities. In the discussion below, we quantify the dependence of the provider's achieved revenue on the fairness threshold ϕ.

In light of our findings that tradeoffs do not exist when users' resource requirements are sufficiently symmetric or asymmetric, we consider an intermediate scenario in which user 1 has similar resource requirements[19] ($r_{11} = r_{12} = r/2$) and user 2 has dissimilar requirements[20] ($r_{21} = \sigma r, r_{22} = (1-\sigma)r$). In this setting, the users' resource requirements can never be perfectly asymmetric for $0 < \sigma < 1$ and can be symmetric only at $\sigma = 1/2$. In other words, for all values of $\sigma \neq \frac{1}{2}$, there will likely be a fairness-revenue tradeoff that our optimization approach can quantify. Without loss of generality, we suppose that $\sigma < 1/2$. We first consider the threshold ϕ values for which the fairness constraint imposed by the regulator is tight at optimality (i.e., imposing the constraint affects the prices chosen by the cloud provider):

[19] In cloud parlance, this is a balanced job, which uses CPU or IO bandwidth in equal proportion.
[20] These jobs are either CPU-intensive or IO-intensive.

Lemma 4. *The fairness constraint in (8) is not tight at optimality if*

$$\phi < \frac{2^{\frac{-(1-\gamma)^2}{\gamma}} c\gamma}{(1-\gamma) r^{1-\gamma} \left((1-\sigma)^{\frac{1}{\gamma}} + (1-\sigma) 2^{1-\frac{1}{\gamma}} \right)^{1-\gamma}}, \tag{9}$$

under which the optimal prices satisfying (8) are given by

$$p_1^{\frac{-1}{\gamma}} = \frac{2\left(1-\sigma\right)^{\frac{1}{\gamma}}}{r c^{\frac{1}{\gamma}} \left((1-\sigma)^{\frac{1}{\gamma}} + (1-\sigma)\, 2^{1-\frac{1}{\gamma}} \right)}, \quad p_2^{\frac{-1}{\gamma}} = \frac{2^{1-\frac{1}{\gamma}}}{r c^{\frac{1}{\gamma}} \left((1-\sigma)^{\frac{1}{\gamma}} + (1-\sigma)\, 2^{1-\frac{1}{\gamma}} \right)}. \tag{10}$$

Moreover, (8) is infeasible if

$$\phi \geq \frac{2^{1-\gamma} c\gamma}{(1-\gamma) r^{1-\gamma} (3-2\sigma)^{1-\gamma}}. \tag{11}$$

Thus, the regulator would only choose thresholds ϕ between the bounds in (9) and (11). Within these bounds, we can now find the achieved revenue as a function of ϕ by first observing that at the optimal prices that solve (8), resource 1's capacity constraint is tight: either resource 1's or resource 2's capacity constraint must be tight at optimality, and since we assume $\sigma \leq 1/2$, resource 2's constraint is always satisfied if resource 1's capacity constraint is satisfied. We can thus write

$$x^\star\left(p_1\right) = \frac{2}{r} - 2(1-\sigma) x^\star\left(p_2\right), \tag{12}$$

due to resource 1's capacity constraint being tight at optimality. This observation allows us to solve for the optimal prices:

Proposition 3. *The optimal prices that solve (8) are given by*

$$p_1^{\frac{-1}{\gamma}} = \frac{2}{r c^{\frac{1}{\gamma}}} - 2(1-\sigma) \left(\frac{\phi(1-\gamma)}{\gamma c^{\frac{1}{\gamma}}} \right)^{\frac{1}{1-\gamma}}, \quad p_2^{\frac{-1}{\gamma}} = \left(\frac{\phi(1-\gamma)}{\gamma c^{\frac{1}{\gamma}}} \right)^{\frac{1}{1-\gamma}} \tag{13}$$

for ϕ satisfying the upper and lower bounds in (9) and (11). Moreover, at these optimal prices $p_1 \geq p_2$. At these prices, the cloud provider's achieved revenue is given by

$$\frac{\phi(1-\gamma)}{\gamma} + 2^{1-\gamma} c \left(\frac{1}{r} - c^{\frac{-1}{1-\gamma}} (1-\sigma) \left(\frac{\phi(1-\gamma)}{\gamma} \right)^{\frac{1}{1-\gamma}} \right)^{1-\gamma} \tag{14}$$

We can see from (13) that as ϕ increases, $p_2^{\frac{-1}{\gamma}}$ also increases, i.e., the two prices have more similar values, as we would expect at a stricter fairness threshold. Moreover, the achieved revenue attains its maximum value at $\sigma = 1/2$, i.e., symmetric users: in this case both users' resource requirements align with the resource requirements, allowing users to process more jobs while still satisfying the resource constraints, and thus allowing the cloud provider to extract more revenue. We can further characterize the achieved optimal revenue as ϕ varies:

Corollary 1. *The optimal revenue (14) decreases as ϕ increases. Moreover, the rate of decrease becomes more negative as ϕ increases (i.e., (14) is a concave function of ϕ).*

Thus, initially increasing ϕ from its minimum value $\underline{\phi}$ has comparatively little effect on the achieved revenue. However, as ϕ increases, the consequent decrease in revenue becomes steeper. A regulator might then wish to limit the threshold ϕ so as to limit the effect on the cloud provider's revenue: there is an especially large risk of unduly harming the provider's revenue when ϕ is close to its maximum value, i.e., a strict fairness threshold.

We finally consider the limits of this tradeoff, for which the fairness threshold ϕ takes its maximum and minimum values (9) and (11). At these threshold values, the cloud provider would either maximize its revenue (for the minimum threshold ϕ) or maximize its fairness (for the maximum threshold ϕ). We can then find the achieved fairness and revenue respectively at these two extremes:

Proposition 4. *Let (p_1^r, p_2^r) denote the revenue-maximizing prices, and $\left(p_1^f, p_2^f\right)$ the fairness-maximizing prices. Then the ratios of the achieved fairness and revenue compared to their*

maximum values are respectively

$$L_f = \frac{\min\left\{U^\star\left(p_1^r\right), U^\star\left(p_2^r\right)\right\}}{\min\left\{U^\star\left(p_1^f\right), U^\star\left(p_2^f\right)\right\}} = \left(\frac{3-2\sigma}{2^{\frac{1}{\gamma}}\left(1-\sigma\right)^{\frac{1}{\gamma}}+2-2\sigma}\right)^{1-\gamma} \tag{15}$$

$$L_r = \frac{p_1^f x^\star\left(p_1^f\right) + p_2^f x^\star\left(p_2^f\right)}{p_1^r x^\star\left(p_1^r\right) + p_2^r x^\star\left(p_2^r\right)} = 2\left(3-2\sigma\right)^{\gamma-1}\left(1+2^{1-\frac{1}{\gamma}}\left(1-\sigma\right)^{1-\frac{1}{\gamma}}\right)^{-\gamma}. \tag{16}$$

Moreover, there is a larger reduction in fairness compared to revenue: $L_f \leq L_r$.

By imposing the constraint that $\min\left\{U^\star\left(p_1\right), U^\star\left(p_2\right)\right\} \geq \phi$ on the cloud provider's prices, the regulator can ensure that the ratio of achieved to maximum fairness is greater than L_f, thus mitigating the loss in fairness due to the cloud provider's attempt to maximize its revenue.

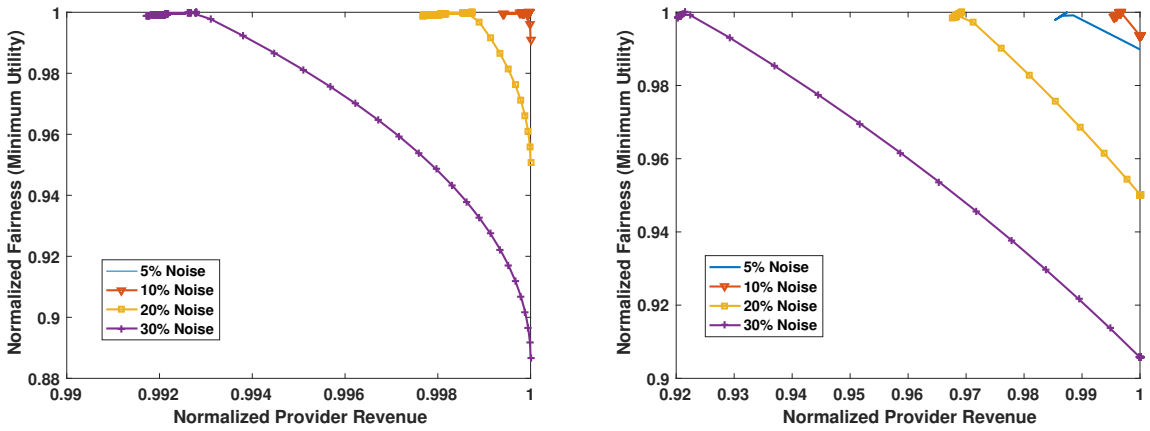
3.4 Numerical Evaluations

In this section, we numerically illustrate the results from Sections 3.2 and 3.3 on the fairness-revenue tradeoff for different configurations of resource requirements among users. Furthermore, we demonstrate the robustness of these results by showing that they remain qualitatively valid for a wide range of scenarios, such as resource allocation decisions with more than two users, choice of different fairness measures (e.g., proportional fairness), different values for the concavity parameter in the users' isoelastic utility functions, and non-isoelastic utility functions (e.g., exponential utility).

We begin our evaluation with Proposition 2, which provides the conditions for the existence of a fairness-revenue tradeoff. In particular, it implies that when the resource requirements for the two users are either sufficiently "symmetric" or perfectly "asymmetric" (as defined in Section 3.2), a significant tradeoff does not arise. But how does this tradeoff change as the resource requirements of the users deviate from the symmetric or asymmetric configurations? To study how the tradeoff evolves numerically, we choose two baseline re-

source configurations (symmetric in Figure 1(a) and asymmetric in Figure 1(b)) for which the tradeoff does not exist, and then perturb these resource requirements r_{ij} by adding independently drawn random noises from uniform distributions with different upper bounds. The resulting fairness-tradeoff plots for four different upper-bounds (5%, 10%, 20%, and 30% over the baseline resource requirement values) are shown in Figure 1. As the percentage of the added noise increases, we observe larger deviations from the initial symmetric or asymmetric resource configuration. For all the numerical plots, we assume that the available capacities of the two resources are normalized to 1 and that the isoelasticity parameter in the users' utility function is $\gamma = 0.33$. The two axes in the plots are normalized with respect to the maximum achievable fairness and revenue values, given the realized resource requirements. When there is no noise (i.e., the configuration is perfectly symmetric or asymmetric), then the cloud provider will be able to maximize both fairness and revenue, i.e., $(1, 1)$ in the figure. As the noise level increases and the resource requirements deviate from the no-tradeoff conditions of Proposition 2, the fairness-revenue tradeoff curves move inwards, i.e., the tradeoff grows more and more severe.

Next, we extend the scenario in Figure 1 to study whether these tradeoffs also arise when there are more than two users. In Figure 2a, we consider perturbations from a scenario in which four users all have symmetric requirements for resource 1, which dominate their requirements for resource 2. As before, there is no fairness-revenue tradeoff with the symmetric requirements, but the tradeoff emerges as we increase the perturbations in resource requirements, thereby decreasing the symmetry in the resource requirements among the four users. We also observe that the tradeoff is more severe for revenue: the loss in fairness when revenue is maximized, $1 - L_f$, exceeds the loss in revenue when fairness is maximized, $1 - L_r$. We also observe this in Figure 2b, which shows the fairness-revenue tradeoff with a baseline of asymmetric resource requirements. Since we have four instead of two users, the notion of perfect resource asymmetry does not easily extend, and so there is a small fairness-revenue



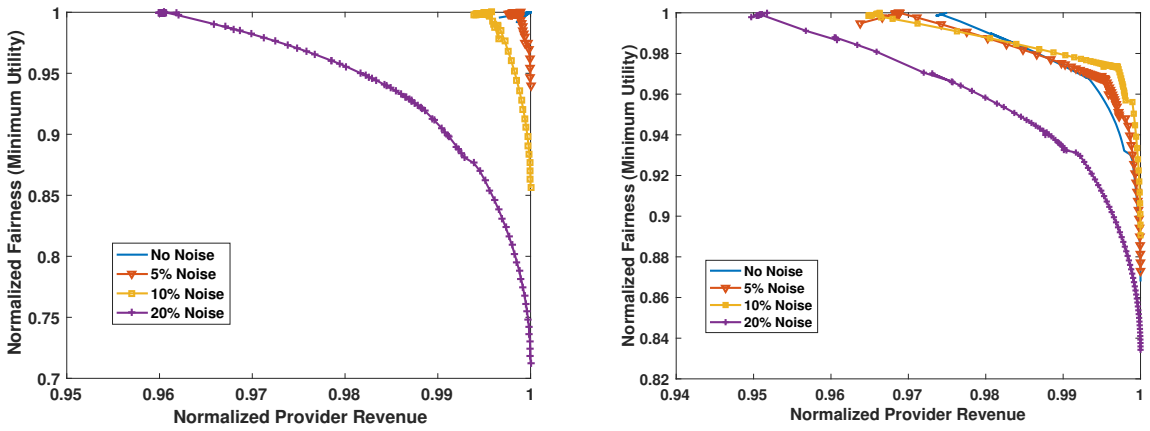
(a) Symmetric requirements.

(b) Asymmetric requirements.

Figure 1: Emergence of (maximin) fairness-revenue tradeoffs when the user's resource requirements deviate from the conditions of Proposition 2. We consider (a) a symmetric resource case with $r_{11} = r_{12} = r_{21} = 0.5$, $r_{22} = 0.33$ and (b) an asymmetric case with $r_{11} = 5/6$, $r_{12} = 1/3$, $r_{21} = 1/2$, $r_{22} = 1$.

tradeoff even without the perturbations. But as the perturbations increase, inducing greater variation in the ratios of the resource requirements among users, the tradeoff becomes more severe, as in the previous case. These results imply that there is a case to be made for a regulatory intervention to ensure that the social objective of fairness, not just the provider's revenue, is accounted for in multi-resource, multi-user settings.

Next, we investigate the fairness-revenue tradeoffs arising from the optimization formulation of Section 3.3 for two users and two resource cases under three different configurations of the resource requirements: two of them are of the form $(r_{11}, r_{12}, r_{21}, r_{22}) = (0.5, 0.5, \sigma, 1-\sigma)$ for $\sigma = 0.2$ and $\sigma = 0.4$, and the third one is derived from the real workload trace of a Google datacenter [22]. For the datacenter trace data, we cluster the jobs according to their CPU and memory requirements and take the resource requirements for each cluster to lie at the cluster centroids to get: $r_{11} = 0.68$, $r_{12} = 0.14$, $r_{21} = 0.32$, $r_{22} = 0.86$. As expected from Proposition 3, we see in Figure 3a that the achieved revenue decreases as a higher fairness threshold is chosen to increase fairness. Moreover, as σ increases (indicating more symmetry



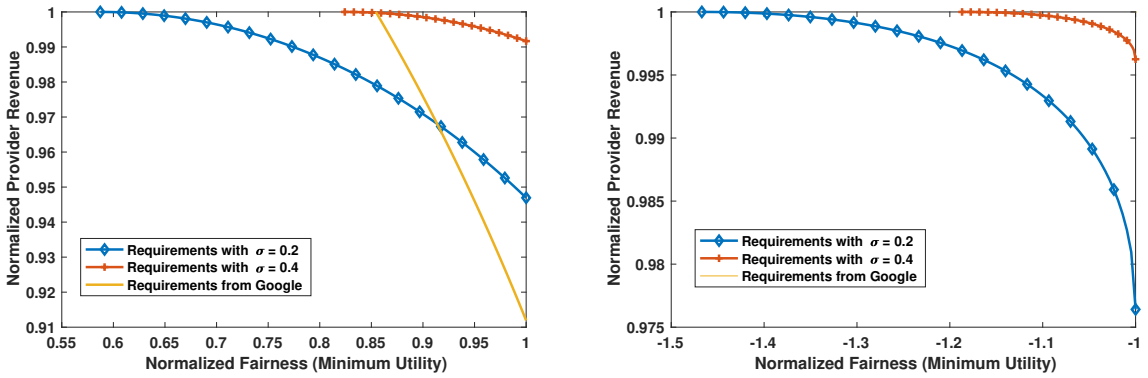
(a) Symmetric requirements.

(b) Asymmetric requirements.

Figure 2: Fairness-revenue tradeoffs for 4 users and 2 resources when the user's resource requirements deviate from symmetric and asymmetric resource requirements. We consider (a) a symmetric resource case with $r_{11} = r_{12} = r_{21} = r_{31} = r_{41} = 0.5$, $r_{22} = r_{32} = r_{42} = 0.33$ and (b) an asymmetric case with $r_{11} = 5/6$, $r_{12} = 1/6$, $r_{21} = 1/2$, $r_{22} = 1/2$, $r_{31} = 2/3$, $r_{32} = 1/3$, $r_{41} = 1/3$, $r_{42} = 1$.

in the two users' resource requirements), the revenue-fairness tradeoff becomes less severe. We also validate the results in Proposition 4, showing that maximizing revenue leads to a greater reduction in fairness (up to 41% when $\sigma = 0.2$) than the reduction in revenue due to maximizing fairness ($< 6\%$). We also note that the requirements from the Google trace lead to a less extreme decrease in fairness when revenue is maximized, but a larger decrease in revenue as the fairness constraint grows tighter. Both of our analytical results, however, still hold: the achieved revenue decreases for a high fairness threshold, and the reduction in fairness from maximizing revenue (15%) is larger than the reduction in revenue from maximizing fairness (9%).

To demonstrate the robustness of these results, we next consider proportional fairness instead of maximin fairness in the optimization model. Figure 3b shows that these results are qualitatively similar when a proportional fairness criterion is used to specify the fairness constraint. The normalized proportional fairness values are negative due to negative values generated by the logarithmic form of the proportional fairness metric. As before, we see that



(a) Max-min fairness.

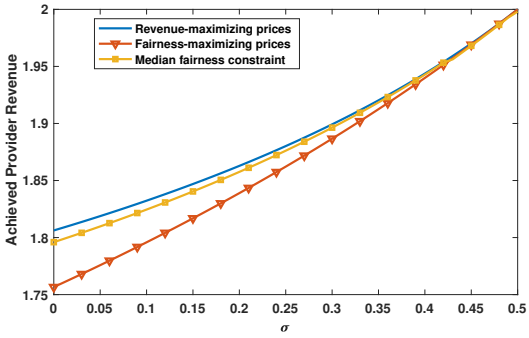
(b) Proportional fairness.

Figure 3: Achieved provider revenue for different fairness thresholds in (8) and different sets of resource requirements. The three sets of resource requirements correspond to $r_{11} = r_{12} = 0.5, r_{21} = \sigma, r_{22} = 1 - \sigma$ for $\sigma = 0.2, 0.4$; and $r_{11} = 0.68, r_{12} = 0.32, r_{21} = 0.14, r_{22} = 0.86$. The last set of resource requirements was derived from the Google datacenter trace. There is no revenue-fairness tradeoff for the Google resource requirements when we use a proportional fairness constraint, and thus the curve is not visible.

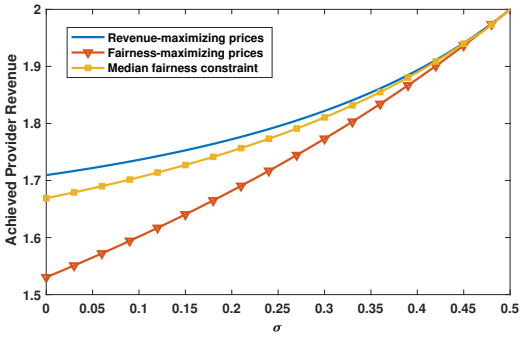
a revenue-fairness tradeoff exists and that it is less severe for larger values of σ.

We now explore the robustness of the results for different value of the parameter γ in the users' isoelastic utility functions. Figure 4 shows the achieved revenue for different resource requirements (i.e., different values of σ) under the revenue-maximizing prices, fairness-maximizing prices, and revenue-maximizing prices subject to a fairness constraint in which the threshold ϕ is set to the midpoint of the maximum and minimum achievable fairness. We see that the achieved revenue always increases with σ for all three types of prices, suggesting that the provider can extract more revenue whenever users' requirements are closer to being symmetric. This finding holds for both the utility function parameters $\gamma = 0.34$ and $\gamma = 0.68$. In both cases, the revenue does not vary too much for the three different sets of prices, which is consistent with Figure 3: a stricter fairness constraint impacts the achieved revenue less than maximizing revenue impacts the achieved fairness.

Lastly, we evaluate the tradeoffs between (maximin) fairness, revenue, and social welfare (i.e., utilitarian fairness) when the users' utility functions are exponential (i.e., specified as



(a) $\alpha = 0.34$.



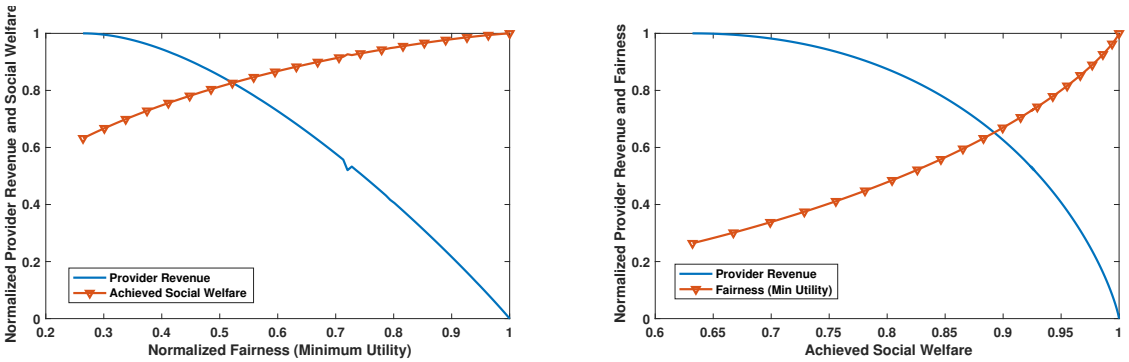
(b) $\alpha = 0.68$.

Figure 4: Achieved provider revenue for different values of α as the resource requirements change. We suppose, as considered in Section 3.3, that users' resource requirements are given by $(r_{11} = r_{12} = r/2)$ and user 2 has dissimilar requirements $(r_{21} = \sigma r, r_{22} = (1-\sigma)r)$. The achieved revenue then depends on the fairness threshold ϕ; we show the achieved revenue when ϕ attains its maximum and minimum values, as well as the midpoint of the two ("median fairness constraint"). The area between the top and bottom curves can be interpreted as the region of achieved revenue for different threshold values.

$V(x) = 1 - \exp(-\lambda x)$ for a given parameter λ) rather than isoelastic. Once again, we use the optimization model in (8) to compute the prices that maximize revenue subject to a constraint on the fairness value. Figure 5 shows our results. We observe that, while there is a similar fairness-revenue tradeoff as found in the isoelastic case, there is also a tradeoff between revenue and social welfare (utilitarian fairness). For isoelastic utility functions, maximizing revenue is equivalent to maximizing the utilitarian social welfare (Lemma 1); but in the exponential utility case, the overall social welfare decreases as revenue increases. However, social welfare increases with the fairness threshold, which means that when users' utility functions are exponential, the maximin fairness constraint imposed by a regulator on the provider's revenue maximization problem (8) will also protect (utilitarian) social welfare.

4 Discussion and Conclusion

In this work, we develop a model for multi-resource allocation in cloud computing and investigate the tradeoffs that different allocations create between fairness and revenue. While



(a) As fairness varies.

(b) As social welfare varies.

Figure 5: Plots of tradeoffs between provider revenue, social welfare, and maximin fairness on user utilities when users have exponential utility functions. Each point shown represents the achieved revenue, social welfare, and max-min fairness when the provider chooses the prices so as to maximize its revenue, subject to a fairness threshold constraint. We observe a similar revenue-fairness tradeoff as with the isoelastic utility functions. Additionally, the achieved social welfare increases as the fairness threshold increases; thus, imposing a fairness constraint also leads to higher (utilitarian) social welfare.

revenue is an important objective for cloud providers, ensuring fairness across different clients is also of critical interest, as observed in our own survey of cloud operators and from the growing interest in fairness in the cloud [16, 24, 28]. In contrast to the existing works on this topic, which have mostly analyzed the operational tradeoffs between fairness and capacity utilization without regard to pricing, we introduce an analytical framework that incorporates an endogenous model of user demand and the cloud providers' pricing and allocation decisions in the multi-resource setting. We discuss the different notions of fairness from welfare economics and then use them in our study to quantify how different configurations of users' resource requirements creates a tradeoff between the fairness and revenue objectives. We show that this tradeoff exists unless users' resource requirements are sufficiently symmetric or sufficiently asymmetric. When a tradeoff exists, a cloud provider's use of revenue-maximizing prices, without consideration of fairness, can lead to a significant reduction in fairness compared to the maximum achievable fairness value. Understanding if and when these tradeoffs arise is required to inform policy decisions, such as whether there is a need for a regulatory intervention to achieve cloud neutrality from the perspective of resource allocation.

The notion of fairness primarily used in the main model is that of maximin fairness, which is useful for a regulator to consider in scenarios where the utility of the least well-off cloud client must be accounted for (e.g., in resource allocation across critical facilities like hospitals and non-profits). Alternatively, the regulator may consider a utilitarian fairness metric in scenarios where equity considerations are less important and some client's jobs are more tolerant of resource starvation (e.g., non-critical, delay-tolerant applications). Proportional fairness may be more relevant to a regulator in scenarios where the overall success of a mission depends on the completion of each client's jobs in proportion to their priorities (e.g., a set of mission-oriented, prioritized military applications [44]). Our numerical investigations demonstrate the robustness of the main results to these alternative fairness specifications.

Our model provides a principled approach for cloud providers to optimize for revenue while requiring a desired level of fairness across its clients to ensure that the transformative benefits of the cloud are available to all. To summarize, in this work (i) the inclusion of social welfare and equity considerations helps motivate a community-level discussion of *what* is fairness, (ii) the identification of the tradeoffs across a wide range of scenarios motivates *why* there is a need for designing new monitoring tools and regulatory policies, (iii) the quantification of resource requirement scenarios that lead to such tradeoffs informs *when* such policy intervention may become necessary, and (iv) the optimization formulation for the cloud provider helps answer *how* such allocations should be made to tradeoff between fairness and revenue goals.

Our work represents an initial foray into developing an analytical framework for considering both social and operational objectives in cloud computing. While we have characterized the case of two users and two resources, deriving analytical results for more general settings is difficult. Our numerical evaluations however show that the findings generally hold for a much wider range of scenarios. Future work can use our framework to investigate, e.g., conditions on multiple users' requirements for which a fairness-revenue tradeoff does or does

not hold. We can also consider incorporating other provider or user objectives, e.g., max-min fairness on the number of jobs submitted by each user instead of user utilities; or the amount of utilized resource capacities instead of provider revenue. Another aspect to further investigate is the role of alternative pricing models, such as pay-per-resource, although we believe that tradeoffs will continue to arise, at least for scenarios where the providers cannot use higher-order price-discrimination across users on each resource. Thus, our framework can be extended for further exploration into a rich set of social and operational tradeoffs in cloud computing. In conclusion, this work can help the IS community to make important policy contributions in the emerging debate on cloud neutrality.

References

[1] E. Altman, K. Avrachenkov, and A. Garnaev. Generalized α-fair resource allocation in wireless networks. In *Proc. of CDC*, pages 2414–2419. IEEE, 2008.

[2] Amazon. EC2 Pricing, September 2016. http://aws.amazon.com/ec2/pricing/.

[3] Anonymous. Harnessing the power of cloud: Revenue and fairness considerations in multi-resource allocations. Technical report, 2017. `https://www.dropbox.com/s/h2lku74urf3lge1/JMIS-main.pdf?dl=0`.

[4] A. B. Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

[5] D. P. Bertsekas and D. El Baz. Distributed asynchronous relaxation methods for convex network flow problems. *SIAM Journal on Control and Optimization*, 25(1):74–85, 1987.

[6] D. Bertsimas, V. F. Farias, and N. Trichakis. The price of fairness. *Operations research*, 59(1):17–31, 2011.

[7] D. Bertsimas, V. F. Farias, and N. Trichakis. On the efficiency-fairness trade-off. *Management Science*, 58(12):2234–2250, 2012.

[8] K. Binmore, A. Rubinstein, and A. Wolinsky. The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, pages 176–188, 1986.

[9] H. K. Cheng, S. Bandyopadhyay, and H. Guo. The debate on net neutrality: A policy perspective. *Information Systems Research*, 22(1):60–82, 2011.

[10] S. Cho, L. Qiu, and S. Bandyopadhyay. Should online content providers be allowed to subsidize content? an economic analysis. *Information Systems Research*, 27(3):580–595, 2016.

[11] L. Columbus. Making cloud computing pay. Forbes, 2013. http://tinyurl.com/csqa9wq.

[12] K. Dean. Cloud and carrier-neutrality in a colocation data centre. Interxion, 2015. `http://www.interxion.com/globalassets/_documents/whitepapers-and-pdfs/datacentres/cloud-and-carrier-neutrality/WP_CLOUDCARRIER_en_0615.pdf`.

[13] C. Donnelly. Lock-in: Using cloud-neutral technology to avoid it. Computer Weekly, 2017. `http://www.computerweekly.com/blog/Ahead-in-the-Clouds/Lock-in-Using-cloud-neutral-technology-avoid-it`.

[14] Eric. What is cloud neutrality, and why does it matter to your business? Root Level Tech, 2016. `https://rootleveltech.com/what-is-cloud-neutrality/`.

[15] G. Feng, S. Garg, R. Buyya, and W. Li. Revenue maximization using adaptive resource provisioning in cloud computing environments. In *Proc. of the ACM/IEEE Conf. on Grid Computing*, pages 192–200. IEEE, 2012.

[16] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In *Proc. of USENIX NSDI*, pages 24–37. USENIX, 2011.

[17] A. Ghodsi, M. Zaharia, S. Shenker, and I. Stoica. Choosy: Max-min fair sharing for datacenter jobs with constraints. In *Proceedings of the 8th ACM European Conference on Computer Systems*, pages 365–378. ACM, 2013.

[18] M. Giacomini, J. Hurley, and D. DeJean. Fair reckoning: a qualitative investigation of responses to an economic health resource allocation survey. *Health Expectations*, 17(2):174–185, 2014.

[19] M. J. K. Gloria. Regulating the cloud: Policy for computing infrastructure, edited by christopher s. yoo and jean-francois blanchette, mit press. *The Communication Review*, 19(2):153–155, 2016.

[20] Google. Compute Engine Pricing, September 2016. `http://bit.ly/2rxGuSi`.

[21] H. Guo, S. Bandyopadhyay, A. Lim, Y.-C. Yang, and H. K. Cheng. Effects of competition among internet service providers and content providers on the net neutrality debate. *MISQ*, 41(2):353–370, 2017.

[22] J. L. Hellerstein. Google cluster data. Google research blog, Jan 2010. `http://bit.ly/2rnec8A`.

[23] IBM. IBM smartcloud workload automation. Data Sheet, 2012. `http://bit.ly/2sr6TQF`.

[24] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang. Multiresource allocation: Fairness-efficiency tradeoffs in a unifying framework. *IEEE/ACM Transactions on Networking*, 21(6):1785–1798, 2013.

[25] K. Weins. Cloud computing trends: 2016 state of the cloud survey. Cloud Management Blog, 2016. `http://bit.ly/1TTf92A`.

[26] E. Kalai and M. Smorodinsky. Other solutions to nash's bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 513–518, 1975.

[27] 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*, 49(3):237–252, 1998.

[28] G. Kesidis, B. Urgaonkar, N. Nasiriani, and C. Wang. Neutrality in future public clouds: Implications and challenges. In *8th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 16)*, 2016.

[29] T. Lan, D. Kao, M. Chiang, and A. Sabharwal. An axiomatic theory of fairness in network resource allocation. In *Proc. of IEEE INFOCOM*, pages 1–9. IEEE, 2010.

[30] S. Lehmann, T. Draisbach, P. Buxmann, and P. Dörsam. Pricing of software as a service–an empirical study in view of the economics of information theory. *Software Business*, pages 1–14, 2012.

[31] H. Li, J. Liu, and G. Tang. A pricing algorithm for cloud computing resources. In *Proc. of the Intl. Conf. on NCIS*, volume 1, pages 69–73. IEEE, 2011.

[32] W.-Y. Lin, G.-Y. Lin, and H.-Y. Wei. Dynamic auction mechanism for cloud resource allocation. In *IEEE Conf. on CCGrid)*, pages 591–592. IEEE, 2010.

[33] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi. Cloud computing: The business perspective. *Decision Support Systems*, 51(1):176–189, 2011.

[34] A. Mas-Colell, M. D. Whinston, J. R. Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

[35] J. Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[36] A. Renda. Competition, neutrality and diversity in the cloud. *Communications and Strategies*, 85:23–44, 2012.

[37] P. Samimi and A. Patel. Review of pricing models for grid & cloud computing. In *Proc. of IEEE ISCI*, pages 634–639. IEEE, 2011.

[38] P. A. Samuelson. Some implications of" linearity.". *The Review of Economic Studies*, 15(2):88–90, 1947.

[39] A. Sen. Utilitarianism and welfarism. *Journal of Philosophy*, 76(9):463–489, 1979.

[40] E. Siham, C. Schlereth, and B. Skiera. Price comparison for infrastructure-as-a-service. In *Proc. of the Eur. Conf. on Information Systems*. AIS, 2012.

[41] A. Tang, D. Wei, and S. H. Low. Heterogeneous congestion control: Efficiency, fairness and design. In *Proc. of ICNP*, pages 127–136. IEEE, 2006.

[42] D. Tilson, K. Lyytinen, and C. Sorensen. Digital infrastructures: The missing IS research agenda. *Information systems research*, 21(4):748–759, 2010.

[43] M. Vizard. The march toward cloud neutrality. Barracuda MSP Industry and Tech Blog, 2017. `https://blog.barracudamsp.com/the-march-toward-cloud-neutrality`.

[44] S. Wagner, E. van den Berg, J. Giacopelli, A. Ghetie, J. Burns, M. Tauil, S. Sen, M. Wang, M. Chiang, T. Lan, R. Laddaga, P. Robertson, and P. Manghwani. Autonomous, collaborative control for resilient cyber defense (ACCORD). In *Workshop on Adaptive Host and Network Security*, 2012.

[45] H. Wang, Q. Jing, R. Chen, B. He, Z. Qian, and L. Zhou. Distributed systems meet economics: Pricing in the cloud. In *Proc. of USENIX HotCloud*, pages 6–6, 2010.

[46] N. Wang, H. Liang, Y. Jia, S. Ge, Y. Xue, and Z. Wang. Cloud computing research in the is discipline: A citation/co-citation analysis. *Decision Support Systems*, 86:35 – 47, 2016.

[47] W. Wang, P. Zhang, T. Lan, and V. Aggarwal. Datacenter net profit optimization with deadline dependent pricing. In *Proc. of CISS*, pages 1–6. IEEE, 2012.

[48] H. Xu and B. Li. Maximizing revenue with dynamic cloud pricing: The infinite horizon case. In *Proc. of IEEE ICC, Next-Generation Networking Symposium*. IEEE, 2012.

[49] 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 *Proc. of the 5th European CCS*, pages 265–278. ACM, 2010.

[50] L. Zheng, C. Joe-Wong, C. W. Tan, M. Chiang, and X. Wang. How to bid the cloud. *ACM SIGCOMM Computer Communication Review*, 45(4):71–84, 2015.

A Proof of Lemma 1

Proof. Substituting $x^\star(p) = c^{\frac{1}{\gamma}} p^{\frac{-1}{\gamma}}$ from (2) into $R(p_1, p_2)$ and $W(p_1, p_2)$ as defined in (4) and (5), we see that $R(p_1, p_2) = c^{\frac{1}{\gamma}} \left(p_1^{\frac{-1}{\gamma}} + p_2^{\frac{-1}{\gamma}} \right)$, $W(p_1, p_2) = \frac{c^{\frac{1}{\gamma}}}{1-\gamma} \left(p_1^{\frac{-1}{\gamma}} + p_2^{\frac{-1}{\gamma}} \right)$. □

B Proof of Lemma 2

Proof. The prices that maximize $\min\{U^\star(p_1), U^\star(p_2)\}$ must solve the maximization problem $\max U^\star(p_i)$, subject to the constraints (3) and $U^\star(p_i) \leq U^\star(p_j)$ for $i \neq j$. We thus wish to choose p_i so as to maximize $U^\star(p_i)$, subject to $p_i \geq p_j$ and (3). Note that if $p_i > p_j$, we could reduce p_i, increasing $U^\star(p_i)$, and increase p_j while maintaining the feasibility of (3); thus, we must have $p_1 = p_2$ at optimality.

Both users then have the same demands $x^\star(p_1) = x^\star(p_2)$. Moreover, $U^\star(p)$ is a decreasing function of p, so the optimal price should be the smallest feasible p. Since $x^\star(p)$ is also

decreasing in p, one of the two resource constraints (3) must be tight at the optimum. We assume that $r_{11} + r_{21} \geq r_{12} + r_{22}$, so the feasibility of resource 1's constraint implies feasibility of resource 2's constraint. Thus, we can assume that $r_{11}x^\star(p) + r_{21}x^\star(p) = 1$, and that $p = c(r_{11} + r_{21})^\gamma$. □

C Proof of Lemma 3

Proof. Let μ, ν be Lagrange multipliers for the resource constraints (3). From the Karush-Kuhn-Tucker conditions, we find necessary conditions for maximizing the revenue $R(p_1, p_2)$ subject to the resource constraints (3):

$$x^\star(p_1) = (\mu r_{11} + \nu r_{12} - p_1)\, x^{\star\prime}(p_1), \quad x^\star(p_2) = (\mu r_{21} + \nu r_{22} - p_2)\, x^{\star\prime}(p_2), \tag{17}$$

which simplifies to (6) using the definition of $x^\star$ from (2). The complementary slackness and feasibility conditions then require that (7) holds and that $\mu, \nu \geq 0$. □

D Proof of Proposition 1

Proof. Consider the fairness-maximizing prices $p_1 = p_2 = c(r_{11} + r_{21})^\gamma$ from Lemma 2. We show that they do not satisfy the condition (7) from Lemma 3 by contradiction. Assuming that these prices also maximize revenue, we can solve for ν, μ in Lemma 3:

$$\begin{bmatrix} \mu \\ \nu \end{bmatrix} = \frac{2c}{r_{11}r_{22} - r_{12}r_{21}} \begin{bmatrix} r_{22} & -r_{12} \\ -r_{21} & r_{11} \end{bmatrix} \begin{bmatrix} (r_{11} + r_{21})^\gamma \\ (r_{11} + r_{21})^\gamma \end{bmatrix}. \tag{18}$$

Adding the constraints (7), we require that $(\mu r_{11} + \nu r_{21})^{1-\frac{1}{\gamma}} + (\mu r_{21} + \nu r_{12})^{1-\frac{1}{\gamma}} = 2^{\frac{-1}{\gamma}} c^{\frac{-1}{\gamma}} (\mu + \nu)$, which from (18) yields the necessary conditions $2(r_{11} + r_{12})^{\gamma-1} = (r_{11} + r_{21})^\gamma \frac{r_{11}+r_{22}-r_{21}-r_{12}}{r_{11}r_{22}-r_{12}r_{21}}$. We then simplify to find the conditions $2 = 1 + \frac{r_{11}^2 - r_{11}r_{12} + r_{21}r_{22} - r_{21}^2}{r_{11}r_{22}-r_{12}r_{21}}$, which upon factorization becomes

$$r_{11}^2 - r_{11}r_{12} + r_{21}r_{22} - r_{21}^2 - r_{11}r_{22} + r_{12}r_{21} = (r_{11} - r_{21})(r_{11} + r_{21} - r_{12} - r_{22}) = 0.$$

By assumption, this equality does not hold. □

E Proof of Proposition 2

Proof. We consider each of the three cases separately, and show that the conditions under which there is a tradeoff for each case fall into one of the two conditions in the proposition.

We first consider the case $r_{11}r_{22} = r_{12}r_{21}$. We then have $r_{11} = r_{12}r_{21}/r_{22}$, and we rewrite the capacity constraint for resource 1 as $r_{21}/r_{22}\left(r_{12}x^\star\left(p_1\right) + r_{22}x^\star\left(p_2\right)\right) \leq 1$. From our assumption $r_{11} \geq r_{12}$, we then have $r_{21} \geq r_{22}$. We then see that resource 2's capacity constraint $r_{12}x^\star\left(p_1\right) + r_{22}x^\star\left(p_2\right) \leq 1$ is always satisfied whenever resource 1's capacity constraint is satisfied and can therefore be ignored in solving the optimization problem. From the Karush-Kuhn-Tucker conditions, we find that at the revenue-maximizing prices, there exists a $\mu > 0$ such that $p_1 = \mu r_{11}/2$, $p_2 = \mu r_{21}/2$. Thus, if $r_{11} = r_{21}$, the revenue-maximizing prices satisfy $p_1 = p_2$, i.e., both users have the same utility, which must be the maximum utility possible under this constraint, and conversely if $r_{11} \neq r_{21}$, then there is a fairness-revenue tradeoff.

We now consider the case $r_{11}r_{22} \neq r_{12}r_{21}$ and $r_{11} = r_{21}$ and show that there is no fairness-revenue tradeoff. We suppose that $r_{11} + r_{21} > r_{12} + r_{22}$, since we will assume that $r_{11} + r_{21} = r_{12} + r_{22}$ in the third case below. Then the KKT conditions for the revenue-maximizing prices allow us to derive $p_1 = \frac{\mu r_{11} + \nu r_{12}}{2}$, $p_2 = \frac{\mu r_{11} + \nu r_{22}}{2}$ for some $\mu, \nu \geq 0$, implying that $p_1 = p_2$, i.e., there is no fairness-revenue tradeoff, if and only if either $r_{12} = r_{22}$ or there does not exist $\nu > 0$ such that the complementary slackness conditions are satisfied and

$$r_{12}c^{\frac{1}{\gamma}}x^\star\left(\frac{\mu r_{11} + \nu r_{12}}{2}\right) + r_{22}c^{\frac{1}{\gamma}}x^\star\left(\frac{\mu r_{11} + \nu r_{22}}{2}\right) = 1. \tag{19}$$

If $r_{12} = r_{22}$, then both users have identical resource requirements, and there is no fairness-revenue tradeoff. In the second case, we suppose that $\nu > 0$ does exist and derive a contradiction. We can assume without loss of generality that $r_{12} > r_{22}$ and find that

$$r_{11}\left(\frac{\mu r_{11} + \nu r_{12}}{2}\right)^{\frac{-1}{\gamma}} + r_{11}\left(\frac{\mu r_{11} + \nu r_{22}}{2}\right)^{\frac{-1}{\gamma}} + (r_{12} - r_{11})\left(\frac{\mu r_{11} + \nu r_{12}}{2}\right)^{\frac{-1}{\gamma}}$$
$$+ (r_{22} - r_{11})\left(\frac{\mu r_{11} + \nu r_{22}}{2}\right)^{\frac{-1}{\gamma}} = c^{\frac{-1}{\gamma}},$$

Thus, since $r_{11}\left(\frac{\mu r_{11}+\nu r_{12}}{2}\right)^{\frac{-1}{\gamma}} + r_{11}\left(\frac{\mu r_{11}+\nu r_{22}}{2}\right)^{\frac{-1}{\gamma}} \leq c^{\frac{-1}{\gamma}}$ from the complementary slackness conditions for resource 1's capacity constraint, we must have

$$(r_{12}-r_{11})\left(\frac{\mu r_{11}+\nu r_{12}}{2}\right)^{\frac{-1}{\gamma}} \geq (r_{11}-r_{22})\left(\frac{\mu r_{11}+\nu r_{22}}{2}\right)^{\frac{-1}{\gamma}},$$

and since $r_{12}-r_{11} < r_{11}-r_{22}$, $r_{22} \geq r_{12}$, which is a contradiction. Thus, there is never a fairness-revenue tradeoff if $r_{11} = r_{21}$.

We finally consider the case $r_{11}r_{22} \neq r_{12}r_{21}$ and $r_{11}+r_{21} = r_{12}+r_{22}$. In this case, the fairness-maximizing prices $p_1 = p_2 = c\,(r_{11}+r_{21})^{\gamma}$ are the unique prices that satisfy the Karush-Kuhn-Tucker necessary conditions for revenue maximization with $\mu, \nu > 0$, i.e., with both resource constraints tight. We now show that another solution satisfying these necessary conditions exists and yields higher revenue if and only if $r_{11} > r_{12} > r_{22}$, i.e., no tradeoff exists if and only if either $r_{11} \geq r_{22} \geq r_{12}$ or $r_{22} \geq r_{11} \geq r_{12}$. Note that we may assume $r_{11} > r_{12}$, as otherwise we would have $r_{11} = r_{12}$ and $r_{21} = r_{22}$. It is then easy to see that a fairness-revenue tradeoff exists unless $r_{11} = r_{12} = r_{21} = r_{22}$.

Suppose that there exists a set of prices (p_1, p_2) satisfying the Karush-Kuhn-Tucker revenue-maximizing conditions with $\mu = 0$, $\nu > 0$, $r_{11} > r_{12}$. Then we may write $p_1 = \nu r_{12}/2, p_2 = \nu r_{22}/2$. Since $\nu > 0$, resource 2's capacity constraint must be tight at this solution $(r_{12}x^{\star}(p_1) + r_{22}x^{\star}(p_2) = 1)$, allowing us to solve for ν: $\nu = \left(\frac{2^{\frac{-1}{\gamma}}c^{\frac{-1}{\gamma}}}{r_{12}^{1-\frac{1}{\gamma}}+r_{22}^{1-\frac{1}{\gamma}}}\right)^{-\gamma}$. We then find conditions under which this solution is feasible, i.e., resource 1's capacity constraint is satisfied:

$$\frac{2^{\frac{-1}{\gamma}}c^{\frac{-1}{\gamma}}}{r_{12}^{1-\frac{1}{\gamma}}+r_{22}^{1-\frac{1}{\gamma}}}\left(\frac{r_{11}r_{12}^{\frac{-1}{\gamma}}}{2^{\frac{-1}{\gamma}}} + \frac{r_{21}r_{22}^{\frac{-1}{\gamma}}}{2^{\frac{-1}{\gamma}}}\right) \leq c^{\frac{-1}{\gamma}} \quad \Longleftrightarrow \quad r_{11}r_{12}^{\frac{-1}{\gamma}} + r_{21}r_{22}^{\frac{-1}{\gamma}} \leq r_{12}^{1-\frac{1}{\gamma}} + r_{22}^{1-\frac{1}{\gamma}}$$
$$\Longleftrightarrow \quad r_{12}^{\frac{-1}{\gamma}}(r_{11}-r_{12}) \leq r_{22}^{\frac{-1}{\gamma}}(r_{22}-r_{21}).$$

Since we assume that $r_{11} > r_{12}$ and $r_{11}+r_{21} = r_{12}+r_{22}$, implying that $r_{11}-r_{12} = r_{22}-r_{21}$,

we additionally require $r_{12} \geq r_{22}$ for this solution to be feasible, i.e., $r_{11} > r_{12} \geq r_{22}$. There is no feasible solution (no fairness-revenue tradeoff) if instead $r_{11} \geq r_{22} \geq r_{12}$ or $r_{22} \geq r_{11} \geq r_{12}$. If such a solution exists, it yields higher revenue: the revenue at the fair solution is $2c^{1-\frac{1}{\gamma}}(r_{21}+r_{22})^{\gamma-1}$, which is smaller than that at the less fair solution if

$$c^{1-\frac{1}{\gamma}}\left(r_{12}^{1-\frac{1}{\gamma}}+r_{22}^{1-\frac{1}{\gamma}}\right)^{\gamma} \geq 2c^{1-\frac{1}{\gamma}}(r_{21}+r_{22})^{\gamma-1} \quad \Longleftrightarrow \quad r_{12}^{1-\frac{1}{\gamma}}+r_{22}^{1-\frac{1}{\gamma}} \geq 2^{\frac{1}{\gamma}}(r_{12}+r_{22})^{1-\frac{1}{\gamma}}$$
$$\Longleftrightarrow \quad \frac{1}{2}\left(r_{12}^{1-\frac{1}{\gamma}}+r_{22}^{1-\frac{1}{\gamma}}\right) \geq \left(\frac{r_{12}}{2}+\frac{r_{22}}{2}\right)^{1-\frac{1}{\gamma}},$$

which is the case since $x^{1-\frac{1}{\gamma}}$ is a convex function.

We now consider the case where a solution exists with $\mu > 0$, $\nu = 0$. In this case, we find that $\mu = \left(\frac{2^{\frac{-1}{\gamma}} c^{\frac{-1}{\gamma}}}{r_{11}^{1-\frac{1}{\gamma}}+r_{12}^{1-\frac{1}{\gamma}}}\right)^{-\gamma}$, and we then have the feasibility condition $r_{22}^{\frac{-1}{\gamma}}(r_{21}-r_{22}) \leq r_{12}^{\frac{-1}{\gamma}}(r_{12}-r_{11})$, which since $r_{11} > r_{12}$ requires that $r_{11} > r_{12} \geq r_{22}$, which are the same conditions as found in the case $\mu = 0, \nu > 0$. Thus, there are no additional conditions under which a fairness-revenue tradeoff exists. □

F Proof of Lemma 4

Proof. We first show that the revenue-maximizing prices that satisfy (8) without any fairness constraint are given by (10), and that these prices yield a fairness value of $F(p_1,p_2) = \frac{2^{\frac{-(1-\gamma)^2}{\gamma}} c\gamma}{(1-\gamma)r^{1-\gamma}\left((1-\sigma)^{\frac{1}{\gamma}}+(1-\sigma)2^{1-\frac{1}{\gamma}}\right)^{1-\gamma}}$. Thus, any fairness threshold less than this value, i.e., satisfying (9), is not tight at optimality.

To show the second part of the lemma, we wish to find the prices (p_1,p_2) that maximize fairness subject to the resource constraints. If ϕ is above this maximum fairness value, which we show is given by (11), then (8) is infeasible. We first note that at the fairness-maximizing prices, we must have $p_1 = p_2$: if $p_1 < p_2$, we could increase p_1 and decrease p_2 so as to increase $\min\{U^\star(p_1), U^\star(p_2)\}$, and vice versa if $p_1 > p_2$. Denoting $p = p_1 = p_2$, we then wish

to find the price p that satisfies

$$\max_{p} \; U^\star(p) \quad \text{s.t. } r\left(\frac{1}{2}+1-\sigma\right)x^\star(p) \le 1,\; p \ge 0. \tag{20}$$

Noting that the constraint $r\left(\frac{1}{2}+1-\sigma\right)x^\star(p) \le 1$ must be tight at optimality, since $U^\star$ is a monotonically increasing function of p, we thus find that $p^{\frac{-1}{\gamma}} = \left(\frac{2c^{\frac{-1}{\gamma}}}{r(3-2\sigma)}\right)^{-\gamma}$ from which we can find the maximum fairness value, $\frac{2^{1-\gamma}c\gamma}{(1-\gamma)r^{1-\gamma}(3-2\sigma)^{1-\gamma}}$. □

G Proof of Proposition 3

Proof. To derive the prices that solve (8), we first note that we can ignore the second resource constraint, $(r/2)x_1^\star(p_1)+r\sigma x_2^\star(p_2) \le 1$, as it is always satisfied if the first resource constraint $(r/2)x_1^\star(p_1)+r(1-\sigma)x_2^\star(p_2) \le 1$ is satisfied. Thus, since this constraint is tight at optimality, we can solve for p_1 in terms of p_2 using (12): $c^{\frac{1}{\gamma}}p_1^{\frac{-1}{\gamma}} = \frac{2}{r} - 2(1-\sigma)c^{\frac{1}{\gamma}}p_2^{\frac{-1}{\gamma}}$. We can then rewrite the optimization problem (8):

$$\begin{aligned} &\max_{p_2} 2^{1-\gamma}\left(\frac{c^{\frac{-1}{\gamma}}}{r} - (1-\sigma)p_2^{\frac{-1}{\gamma}}\right)^{1-\gamma} + p_2^{1-\frac{1}{\gamma}} \\ &\text{s.t. } \frac{\gamma c^{\frac{1}{\gamma}}}{1-\gamma}p_2^{1-\frac{1}{\gamma}} \ge \phi,\; 2^{1-\gamma}\left(\frac{c^{\frac{-1}{\gamma}}}{r} - (1-\sigma)p_2^{\frac{-1}{\gamma}}\right)^{1-\gamma} \ge \phi. \end{aligned} \tag{21}$$

We now take the first derivative of this revenue function (21) with respect to p_2 and find that it is maximized when

$$p_2 = 2^{1-\gamma}(1-\sigma)\left(\frac{c^{\frac{-1}{\gamma}}}{r} - (1-\sigma)p_2^{\frac{-1}{\gamma}}\right)^{-\gamma} p_2^{\frac{-1}{\gamma}-1}, \quad p_2^{\frac{-1}{\gamma}} = \frac{2^{1-\frac{1}{\gamma}}}{rc^{\frac{1}{\gamma}}(1-\sigma)^{\frac{1}{\gamma}} + (1-\sigma)2^{1-\frac{1}{\gamma}}},$$

and we use $\overline{p_2}$ to denote the solution to this equation. Moreover, we find that the derivative is negative if $p_2^{\frac{-1}{\gamma}} > \overline{p_2}^{\frac{-1}{\gamma}}$ and positive otherwise. We now note that, if the revenue-maximizing $p_2 = \overline{p_2}$ does not satisfy the fairness threshold constraint, then we must decrease p_2 until it is satisfied. Since the derivative of revenue with respect to p_2 is positive for $p_2 < \overline{p_2}$, we wish to decrease p_2 as little as possible, i.e., take the maximum p_2 such that $U^\star(p_2) \ge \phi$ is

satisfied. We then have the solution $p_2^{\frac{-1}{\gamma}} = \left(\frac{\phi(1-\gamma)}{\gamma c^{\frac{1}{\gamma}}}\right)^{\frac{1}{1-\gamma}}$ as in (13). To show that this value of p_2 in fact solves (8), we now solve for p_1, yielding the solution in (13) and show that it satisfies the constraints $p_1 \geq 0$, $U^\star(p_1) \geq \phi$. We find that $p_1 \geq 0$ if

$$\frac{2}{rc^{\frac{1}{\gamma}}} \geq 2(1-\sigma)\left(\frac{\phi(1-\gamma)}{\gamma c^{\frac{1}{\gamma}}}\right)^{\frac{1}{1-\gamma}} \iff \phi \leq \frac{\gamma c}{r^{1-\gamma}(1-\gamma)(1-\sigma)^{1-\gamma}},$$

which is always the case if $\phi \leq \frac{2^{1-\gamma}c\gamma}{(1-\gamma)r^{1-\gamma}(3-2\sigma)^{1-\gamma}}$ as in (11):

$$\frac{2^{1-\gamma}c\gamma}{(1-\gamma)r^{1-\gamma}(3-2\sigma)^{1-\gamma}} \leq \frac{\gamma c}{r^{1-\gamma}(1-\gamma)(1-\sigma)^{1-\gamma}} \iff 2(1-\sigma) \leq 3-2\sigma.$$

To check that $U^\star(p_1) \geq \phi$, it suffices to show that $p_1 \leq p_2$:

$$p_1^{\frac{-1}{\gamma}} = \frac{2}{rc^{\frac{1}{\gamma}}} - 2(1-\sigma)p_2^{\frac{-1}{\gamma}} \geq p_2^{\frac{-1}{\gamma}} \iff p_2^{\frac{-1}{\gamma}} \leq \frac{2}{rc^{\frac{1}{\gamma}}(3-2\sigma)},$$

which is equivalent to

$$\left(\frac{\phi(1-\gamma)}{\gamma c^{\frac{1}{\gamma}}}\right)^{\frac{1}{1-\gamma}} \leq \frac{2}{rc^{\frac{1}{\gamma}}(1+2(1-\sigma))} \iff \phi \leq \frac{2^{1-\gamma}\gamma c}{r^{1-\gamma}(1-\gamma)(3-2\sigma)^{1-\gamma}},$$

which holds from Lemma 4. Given these expressions for $p_1^\star$ and $p_2^\star$ solving (13), we can find the revenue (14). □

H Proof of Corollary 1

Proof. We take the derivative of the optimal revenue (14) to find that

$$\frac{dR(p_1^\star, p_2^\star)}{d\phi} = \left(\frac{1}{\gamma} - 1\right)\left(1 - 2^{1-\gamma}c^{\frac{-\gamma}{1-\gamma}}(1-\sigma)\left(\frac{\phi(1-\gamma)}{\gamma}\right)^{\frac{-\gamma}{1-\gamma}}\left(\frac{1}{r} - c^{\frac{-1}{1-\gamma}}(1-\sigma)\left(\frac{\phi(1-\gamma)}{\gamma}\right)^{\frac{1}{1-\gamma}}\right)\right). \tag{22}$$

This derivative is negative if and only if

$$2^{1-\gamma}c^{\frac{-\gamma}{1-\gamma}}(1-\sigma)\left(\frac{\phi(1-\gamma)}{\gamma}\right)^{\frac{-\gamma}{1-\gamma}}\left(\frac{1}{r}-c^{\frac{-1}{1-\gamma}}(1-\sigma)\left(\frac{\phi(1-\gamma)}{\gamma}\right)^{\frac{1}{1-\gamma}}\right)\geq 1$$

$$\frac{1}{r}-c^{\frac{-1}{1-\gamma}}(1-\sigma)\left(\frac{\phi(1-\gamma)}{\gamma}\right)^{\frac{1}{1-\gamma}}\leq 2^{\frac{1}{\gamma}-1}c^{\frac{-1}{1-\gamma}}(1-\sigma)^{\frac{1}{\gamma}}\left(\frac{\phi(1-\gamma)}{\gamma}\right)^{\frac{1}{1-\gamma}}$$

$$\phi\geq\frac{2^{\frac{-(1-\gamma)^2}{\gamma}}c\gamma}{(1-\gamma)r^{1-\gamma}\left((1-\sigma)^{\frac{1}{\gamma}}+(1-\sigma)2^{1-\frac{1}{\gamma}}\right)^{1-\gamma}},$$

which is the minimum practical value of ϕ as derived in Lemma 4; for ϕ below this threshold, $dR/d\phi = 0$ since the fairness threshold is not tight at optimality. By inspection, (22) becomes more negative as ϕ increases. □

I Proof of Proposition 4

Proof. We first find (p_1^r, p_2^r) and $\left(p_1^f, p_2^f\right)$ from the proof of Proposition 3:

$$p_1^f = p_2^f = \left(\frac{2c^{\frac{-1}{\gamma}}}{r\left(3-2\sigma\right)}\right)^{-\gamma}$$

$$p_1^r = \left(\frac{2}{rc^{\frac{1}{\gamma}}\left(2^{1-\frac{1}{\gamma}}\left(1-\sigma\right)^{1-\frac{1}{\gamma}}+1\right)}\right)^{-\gamma}, p_2^r \qquad = \left(\frac{2^{1-\frac{1}{\gamma}}\left(1-\sigma\right)^{\frac{-1}{\gamma}}}{rc^{\frac{1}{\gamma}}\left(2^{1-\frac{1}{\gamma}}\left(1-\sigma\right)^{1-\frac{1}{\gamma}}+1\right)}\right)^{-\gamma}$$

Substituting these prices into the fairness and revenue expressions, we find the percentage losses in (15) and (16). We then show that the percentage loss in fairness is always larger than that in revenue by showing that $L_f/L_r \leq 1$:

$$\frac{L_f}{L_r} = \frac{(3-2\sigma)^{1-\gamma}\left(1+2^{1-\frac{1}{\gamma}}\left(1-\sigma\right)^{1-\frac{1}{\gamma}}\right)^{\gamma}}{2\left(3-2\sigma\right)^{\gamma-1}\left(2-2\sigma+2^{\frac{1}{\gamma}}\left(1-\sigma\right)^{\frac{1}{\gamma}}\right)^{1-\gamma}} = \left(\frac{3-2\sigma}{2-2\sigma}\right)^{2-2\gamma}\frac{\left(1+(2-2\sigma)^{\frac{1}{\gamma}-1}\right)^{2\gamma-1}}{2}. \quad (23)$$

Taking the derivative of this ratio with respect to γ, we find that

$$\frac{d}{d\gamma}\left(\log\left(\frac{L_f}{L_r}\right)\right) = -2\left(\log\left(3-2\sigma\right)-\log\left(2-2\sigma\right)\right)+2\log\left(1+\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}\right)$$
$$-\frac{\left(2\gamma-1\right)\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}\log\left(2-2\sigma\right)}{\gamma^2\left(1+\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}\right)}$$
$$=2\left(\log\left(1+\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}\right)-\log\left(3-2\sigma\right)\right)$$
$$+\log\left(2-2\sigma\right)\left(2-\frac{\left(2\gamma-1\right)\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}}{\gamma^2\left(1+\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}\right)}\right). \tag{24}$$

Note that $1+\left(2-2\sigma\right)^{\frac{1}{\gamma}-1} \geq 3-2\sigma$ when $\gamma \leq 1/2$; thus, to show that this derivative is positive, it suffices to show that $\frac{(2\gamma-1)(2-2\sigma)^{\frac{1}{\gamma}-1}}{\gamma^2\left(1+(2-2\sigma)^{\frac{1}{\gamma}-1}\right)} \leq 2$. Simplifying, we find the equivalent condition $4\gamma^2-2\gamma+1>0$, which is always the case. Thus, it suffices to show that $L_f/L_r \leq 1$ at $\gamma=1$; by inspection, $L_f/L_r=1$ at $\gamma=1$.

We now consider the case $\gamma \geq 1/2$. In this case, we can show that (24) is positive by finding the equivalent condition

$$2\log 2-2\log\left(3-2\sigma\right)+\log\left(2-2\sigma\right)\frac{4\gamma^2-2\gamma+1}{\gamma^2\left(1+\left(2-2\sigma\right)^{\frac{1}{\gamma}-1}\right)}$$
$$>2\left(\log 2-\log\left(3-2\sigma\right)\right)+\frac{\log\left(2-2\sigma\right)}{3-2\sigma}\left(4-\frac{2}{\gamma}+\frac{1}{\gamma^2}\right).$$

Since $4-\frac{2}{\gamma}+\frac{1}{\gamma^2}=\left(\frac{1}{\gamma}-1\right)^2+3 \geq 3$, we find the sufficient condition $2\left(2-2\sigma\right)^{\frac{3}{3-2\sigma}}/(3-2\sigma) > 1$, which is true by inspection at $\sigma=0,1$. We then show that $\log\left(2\left(2-2\sigma\right)^{\frac{3}{3-2\sigma}}/(3-2\sigma)\right)$ is a concave function, i.e., that it is sufficient to check its positivity at $\sigma=0,1$:

$$\frac{d}{d\sigma}\log\left(\frac{2\left(2-2\sigma\right)^{\frac{3}{3-2\sigma}}}{3-2\sigma}\right)=\frac{2}{\log(3-2\sigma)}-\frac{2}{(3-2\sigma)\log(2-2\sigma)}+\frac{2\log(2-2\sigma)}{(3-2\sigma)^2}$$
$$\frac{d^2}{d\sigma^2}\log\left(\frac{2\left(2-2\sigma\right)^{\frac{3}{3-2\sigma}}}{3-2\sigma}\right)=\frac{4}{(3-2\sigma)\log(3-2\sigma)^2}+\frac{2\left(-2\log(2-2\sigma)-2\frac{3-2\sigma}{2-2\sigma}\right)}{(3-2\sigma)^2\log(2-2\sigma)^2}$$
$$+\frac{8\log(2-2\sigma)-2\frac{3-2\sigma}{2-2\sigma}}{(3-2\sigma)^3}.$$

Thus, to show concavity it suffices to show that $8\log(2-2\sigma)<2.5<3<2\left(3-2\sigma\right)/\left(2-2\sigma\right)$. □