Pricing the Cloud: Resource Allocations, Fairness, and Revenue

Carlee Joe-Wong* and Soumya Sen†

*Princeton University and †University of Minnesota

Emails: cjoe@princeton.edu, ssen@umn.edu

Abstract

As more businesses use the cloud for their computing needs, datacenter operators are increasingly pressed to perform effective and fair allocation in this *multi-resource, multi-tenant* setting. The presence of multiple resources allows an operator to offer different types of pricing strategies (e.g., bundled vs. unbundled) that can have different effects on its revenue. Pricing also affects the demand and resource allocation decisions across clients who typically require different ratios of each resource (e.g., CPUs, memory, bandwidth) to process their jobs, which results in a complex trade-off between fairness and revenue maximization. We develop an analytical framework to investigate the fairness and revenue tradeoffs that arise in a datacenter's multi-resource setting and the impact of different pricing plans on the operator's objective. We derive analytical bounds on the operator's fairness-revenue tradeoff and compare tradeoff points for different pricing strategies on a data trace taken from a Google cluster.

1 Introduction

Today's network technologies increasingly allow for better control and sharing of resources across multiple clients, resulting in a distinct trend in the creation of large centralized datacenter facilities. Investment in such datacenters is growing rapidly; in the second quarter of 2013 alone, Google and Microsoft spent $3.4 billion on improving their cloud infrastructure [4]. Recent studies found that businesses realize an average of 22% savings due to cloud computing, with benefits of up to 48% in areas like innovation [2]. By 2020, 40% of digital business information are predicted to use cloud facilities [6]. While some IS studies have compared cloud pricing plans offered by specific operators [9] and client preferences for cloud services [8], there is a need to address several business and economic challenges, in particular on how to allocate and price datacenter resources.

These questions are difficult to answer due to unique characteristics of the cloud computing setting. The cloud has (a) **multiple resources**: data centers need to allocate multiple types of resources (e.g., CPU cycles, memory, bandwidth) to process each job from its clients in a centralized facility with finite capacity, (b) **multiple tenants**: data centers cater to clients whose jobs are heterogeneous in the ratio of each type of resources required, which introduces a complex trade-off between *fairness* in multi-resource allocation and the provider's objective of efficient utilization of these finite resources for *revenue maximization*; and (c) **multiple pricing schemes:** the choice of pricing plan used (e.g., individual resource pricing vs. bundled pricing), the price points, and volume discounts offered, will not only affect client demand, but also the fairness, resource utilization, and revenues resulting from the optimal allocation of these multiple types of resource across clients. Therefore, it is important to develop a systematic approach to study the interplay between pricing, fairness of multi-resource allocation, and revenue of cloud operators.

As an initial step in exploring these issues, we consider an Infrastructure-as-a-Service (IaaS) view of the cloud with non-substitutable computing resources such as CPU time, memory, and

bandwidth, that are shared by multiple clients (tenants) with different ratios of resources required for their jobs. We also consider that the datacenter operator wishes to maximize its revenue while accounting for fairness (i.e., equitability in the number of jobs processed for each client) to avoid incurring client dissatisfaction and churn in this competitive market [5]. Our framework allows operators to balance their revenue and fairness through setting prices in two popular schemes:

Bundled pricing: Bundled pricing assumes that operators sell groups of heterogeneous resources as bundles to clients. For instance, Google's Compute Engine sells virtual machines (VMs), each with a fixed CPU and memory capacity and charges at a unit rate per VM. Depending on the nature of the job (e.g., computationally-intensive versus memory-intensive), clients may or may not be able to make use of all or part of each type of VM resource sold in that bundle.

Resource pricing: Under resource pricing, an operator charges clients separate unit prices for each resource that they use. Many operators, such as Google and Microsoft's Windows Azure, charge such prices for egress bandwidth. Dimension Data, a smaller cloud operator, goes further, charging separate unit prices for client-specified amounts of CPU, RAM, storage, and bandwidth.

Other pricing measures such as dynamic auction are also sometimes offered (e.g., Amazon EC2's spot pricing) but they have been shown to be less effective in practice [10], and are hence not considered in this initial work.

Some datacenters like Amazon's EC2 take advantage of their economies of scale by introducing *volume discounts* into a pricing policy, charging a lower unit rate for clients who purchase a large quantity of resources. While some works have investigated the effects of volume discounts, e.g., for contract clients [3], none have done so in a datacenter's multi-resource context. In this work, we compare bundled and resource pricing with volume discounts in a mathematical framework that takes into account client demand, operator revenue, and fairness across clients. While tradeoffs between fairness and efficiency have recently been considered in the general context of resource sharing [1], these works do not explicitly generalize to the setting of multiple, non-substitutable resources. Works that do consider multi-resource fairness, on the other hand, have not considered the pricing aspects [7]. Our work is unique in developing a generic model for a multi-resource, multi-tenant setting under different pricing schemes.

We begin developing our framework in Section 2 by examining client choices of how many jobs to submit, given a set of offered prices. Section 3 then discusses the operator objectives of revenue and fairness and tradeoffs between these two functions. We compare the fairness and revenue achieved by bundled and resource pricing in Section 4. Finally, in Section 5 we perform simulations on a dataset from a Google cluster to compare the achieved fairness and efficiency for different pricing plans and a range of volume discounts. We conclude the paper in Section 6.

2 Client Optimization

A datacenter operator has multiple (heterogeneous) clients, each of whom submits jobs, each requiring a certain amount of the different types of non-substitutable datacenter resources (CPU time, memory, and bandwidth). We index the resources with the variable i, and denote the amount of resource i required for each job by R_{ij}, where j indexes the client type. We fix the number of types as n and the number of resources as m.

When deciding how many jobs to submit, clients take into account the prices set by the datacenter operator, which yield a fixed cost per job r_j for each client j. We consider two ways to structure this pricing plan: bundled and resource pricing. In both cases, r_j accounts for both the resource

prices and the volume discounts, which are parameterized with a constant $\gamma \in (0,1]$. If a client requires $R_{ij}x_j$ amount of resource i, where x_j denotes the number of jobs submitted by client j, then under resource pricing she is charged for $R_{ij}^{\gamma}x_j^{\gamma}$ amount of resource i. Under bundled pricing, we have a similar expression with bundles treated as one resource. Thus, a larger γ corresponds to a smaller volume discount, and $\gamma = 1$ corresponds to no volume discount. Since the amount of resources required is an affine function of the client's demand, we let r_j denote the per-job cost with the volume discount and separately account for the discounted x_j^{γ} term.

Clients derive utility from the number of jobs that the datacenter processes and pay for the cost of jobs.[1] We use x_j to denote the number of jobs processed for each client j and $U_j(x_j)$ to denote the utility from these jobs. We assume that U_j is a strictly increasing, concave function expressed in units of dollars, e.g., the often-used *isoelastic* functions

$$U_j(x) = \begin{cases} c_j(1-\alpha_j)^{-1}x^{1-\alpha_j} - r_j x_j^{\star\gamma}, & \alpha_j \in (0,1) \\ c_j \log(x+1) - r_j x_j^{\star\gamma} & \alpha_j = 1 \end{cases} \tag{1}$$

where α_j parameterizes the concavity of the utility function (i.e., the degree of diminishing marginal utility) and the constant $c_j > 0$ scales the utility level. As α_j increases, the client becomes more price-sensitive since the cost term $r_j x_j^{\star\gamma}$ term becomes more dominant. We assume that the client receives zero utility if no jobs are processed, i.e., $U_j(0) = 0$.

Each client j maximizes its utility (1). For simplicity, we assume that U_j is continuously differentiable for $x_j > 0$. We suppose that the clients have no explicit budget constraints and pay for any number of jobs x_j, so long as $U_j > 0$; for instance, enterprise clients would meet this requirement. Clients will submit as many jobs as required to maximize U_j for a given set of prices. The operator chooses a value of γ such that there exists x_j^* maximizing $U_j(x_j)$; for instance, the isoelastic utility functions satisfy this condition if $\gamma > 1 - \alpha_j$. Each client's utility at the optimum then equals the amount paid $r_j x_j^{\star\gamma}$, multiplied by a price-independent constant. Given a set of prices $\mathbf{p}$, we have $\overline{U}_j = \left(\gamma/(1-\alpha_j) - 1\right) r_j x_j^{\star}(\mathbf{p})^{\gamma} = \left(\gamma/c_j\right)^{\gamma/(1-\alpha_j-\gamma)} \left(\gamma/(1-\alpha_j) - 1\right) r_j^{(1-\alpha_j)/(1-\alpha_j-\gamma)}$ where $\overline{U}_j$ denotes the client's utility at the utility-maximizing prices.

3 Fairness and Revenue Tradeoffs

Using the notation of Section 2, the operator's revenue may be expressed as $\rho(\mathbf{p}) = \sum_{j=1}^{n} r_j x_j^{*\gamma}(\mathbf{p})$ for a set of bundled or resource prices $\mathbf{p}$. Since clients are heterogeneous in the amount of different resources needed by their jobs, a simple profit maximization by an operator could inadvertently favor one client over another, possibly introducing client dissatisfaction. For instance, suppose that one client k requires a relatively large amount of one resource l (e.g., this client runs a processor-intensive task like video encoding). Under resource pricing, the operator might increase the price p_l of this resource significantly, in order to offer much lower prices for the other resources. But client k would then have a higher per-job cost $\sum_{i=1}^{m} R_{ik}^{\gamma} p_i$ than other clients, which would keep demand for each resource under capacity but yield lower utility U_k for client k. To avoid this scenario, the operator can consider the fairness of client utilities.

Since the operator's choice of prices results in resources being allocated to each client, this decision can be modeled in the *multi-resource allocation* framework of [7]. We modify this

[1] We assume throughout that the operator sets prices so that all submitted jobs can be processed with the resources available. "Submitted" and "processed" can thus be used interchangeably.

approach in our framework to measure the fairness of the surplus $\overline{U}_j = U_j\left(x_j^\star\right)$ received by each client j. Following the fairness measures proposed in [7], we choose the fairness functions $F_\beta(\mathbf{p}) = \frac{1}{1-\beta}\sum_{j=1}^n \overline{U}_j^{1-\beta}$, where $\beta > 0$ parameterizes the type of fairness. This fairness function corresponds to the well-known α-*fairness function* with $\alpha = \beta$; we thus observe that as β grows, F_β becomes "more fair" [7].

In order to optimize both revenue and fairness, we suppose that operators consider a weighted sum $\nu\rho(\mathbf{p}) + F_\beta(\mathbf{p})$ of the revenue ρ and fairness F_β, where $\nu > 0$ parameterizes the operator's relative emphasis on revenue over fairness.[2] We next examine the range of the tradeoff between fairness and revenue for bundled and resource pricing. For bundled pricing, there is no tradeoff:

Proposition 1. *Under bundled pricing, there is no tradeoff between fairness and revenue.*

Proof. Bundled pricing may be viewed as a single-resource scenario: under bundled pricing, one bundle of resources at fixed ratios may be regarded as the only "resource" offered to clients. The resource capacity C is then the number of bundles, and each client's resource requirement R_j is the number of bundles required to complete one job. We thus prove the proposition when only one resource is offered. The lowest feasible price for this resource is the price for which the resource constraint $\sum_{j=1}^n R_j x_j^* \le C$ is tight, where the left hand side equals the amount of resource demanded by all clients. Since the demands $x_j^\star$ are decreasing in the resource price, a unique price exists satisfying this property. If clients have isoelastic utilities, then we can differentiate both the fairness and revenue expressions with respect to the price to show that the lowest feasible price maximizes both fairness and revenue. There is then no tradeoff between fairness and revenue. □

Under resource pricing, a tradeoff exists, though we can lower-bound the worst-case outcome:

Proposition 2. *At any set of per-job prices r_j, achieved fairness can be lower-bounded by achieved revenue and vice versa. If $\beta > 1$, the revenue ρ can be lower-bounded in terms of fairness:*

$$\rho(\mathbf{p}) \ge \left(F_\beta(\mathbf{p})(1-\beta)\right)^{\frac{1}{1-\beta}} \sum_{j=1}^n \frac{1-\alpha_j}{\gamma+\alpha_j-1}.$$

If $\beta < 1$, then the fairness $F_\beta(\mathbf{p})$ may be lower-bounded in terms of the achieved revenue:

$$F_\beta(\mathbf{p}) \ge \frac{\rho(\mathbf{p})^{1-\beta}}{1-\beta}\left(\frac{\gamma}{1-\alpha_k}-1\right)^{1-\beta}, \quad \alpha_k = \max_j \alpha_j.$$

Proof. Suppose that $\beta < 1$. We first introduce the notation $\phi_j = \frac{\gamma+\alpha_j-1}{1-\alpha_j} = \frac{\gamma}{1-\alpha_j} - 1$, and find that the fairness function may be written as

$$\begin{aligned}\frac{1}{1-\beta}\sum_{j=1}^n \phi_j^{1-\beta}\left(\frac{\gamma}{c_j}\right)^{\frac{\gamma(1-\beta)}{1-\alpha_j-\gamma}} r_j^{\frac{(1-\alpha_j)(1-\beta)}{1-\alpha_j-\gamma}} &\ge \frac{\min_j \phi_j^{1-\beta}}{1-\beta}\left(\sum_{j=1}^n \left(\frac{\gamma}{c_j}\right)^{\frac{\gamma}{1-\alpha_j-\gamma}} r_j^{\frac{1-\alpha_j}{1-\alpha_j-\gamma}}\right)^{1-\beta} \\ &= \frac{\rho(\mathbf{p})^{1-\beta}}{1-\beta}\left(\frac{\gamma}{1-\alpha_k}-1\right)^{1-\beta}.\end{aligned}$$

[2] In some cases, maximizing $\nu\rho(\mathbf{p}) + F_\beta(\mathbf{p})$ is particularly easy numerically; we can differentiate $\nu\rho + F_\beta$ to show that, when the revenue weight ν is sufficiently small, maximizing this objective is a convex optimization that can be solved with standard algorithms.

The first inequality uses the sub-additivity of the function $f(x) = x^{1-\beta}$, and we let $k = \text{argmax}_j \alpha_j$.

We now take $\beta > 1$ and find that for each client j, $\phi_j^{1-\beta} \left(\gamma/c_j\right)^{\frac{\gamma(1-\beta)}{1-\alpha_j-\gamma}} r_j^{\frac{(1-\alpha_j)(1-\beta)}{1-\alpha_j-\gamma}} \leq F_\beta(\mathbf{p})(1-\beta)$, which is equivalent to $\left(\gamma/c_j\right)^{\frac{\gamma}{1-\alpha_j-\gamma}} r_j^{\frac{1-\alpha_j}{1-\alpha_j-\gamma}} \geq \phi_j^{-1} \left(F_\beta(\mathbf{p})(1-\beta)\right)^{\frac{1}{1-\beta}}$. Adding these lower bounds and using the definition of ν_j, we see that $\rho(\mathbf{p}) \geq \left(F_\beta(\mathbf{p})(1-\beta)\right)^{\frac{1}{1-\beta}} \sum_{j=1}^n \frac{1-\alpha_j}{\gamma+\alpha_j-1}$. □

In Section 5, we numerically investigate the fairness-revenue tradeoff for different volume discounts and distributions of users. We found that for a wide range of parameter values, resource pricing yields only a limited tradeoff; the achieved fairness (revenue) when revenue (fairness) is maximized is not far from the maximum fairness (revenue) under resource pricing.

In addition to this fairness-revenue tradeoff, the operator's pricing plan will affect the achieved fairness and revenue. We turn to this topic in the next section.

4 Pricing Datacenter Jobs

We first discuss bundled and resource pricing in Section 4.1 and then compare the optimal prices and resulting revenue and fairness in Section 4.2.

4.1 Pricing Strategies

We first consider **bundled pricing**, in which the operator groups different resources in a fixed ratio–for instance, 1 CPU and 2 GB of RAM might be grouped together as a bundled resource. We consider a single bundle option, and personalized bundles can be mapped to a resource pricing case. The operator then charges a unit price per bundle of resources required for a client's jobs. We let b_i denote the amount of each resource $i = 1, \ldots, m$ in each bundle, and let p denote the per-bundle price. Using the notation from Section 2, we find that each client j requires $\mu_j = \max_i \left(R_{ij}/b_i\right)$ bundles in order to complete one job. We incorporate the volume discount to find the per-job cost $r_j = \mu_j^\gamma p$. Moreover, the number of bundles available is $\min_i C_i/b_i$, where C_i denotes the capacity of resource i. Thus, the operator faces the resource constraint $\sum_{j=1}^n \max_i \left(R_{ij}/b_i\right) x_j^\star \left(\mu_j^\gamma p\right) \leq \min_i \left(C_i/b_i\right)$, where $x_j^\star \left(\mu_j^\gamma p\right)$ denotes the client's demand for jobs at price p, given the bundle $(b_1, \ldots, b_m)$. The resulting revenue under bundled pricing, ρ_b, may be expressed as $\rho_b = p \sum_{j=1}^n \left(\mu_j x_j^\star \left(\mu_j^\gamma p\right)\right)^\gamma$.

Since ρ_b, like any function of the per-job prices $r_j = \mu_j^\gamma p$, is a non-differentiable function of the amounts b_i of resource i in the bundle, the operator may find it computationally difficult to optimize over the b_i. Instead, it can fix b_i, the amount of resource i in each bundle, and optimize over the unit price p. For instance, the b_i might be chosen proportional to the resource capacities C_i, i.e., $b_i = bC_i$ for all i and a constant b. In this case, the client's bundle requirement μ_j is proportional to her **dominant share**, defined as the maximum share of any resource allocated to the client.[3]

To avoid forcing clients to purchase resources they may not need, an operator could instead offer **resource pricing**, in which a unit price is charged for each resource and clients may purchase exactly as much of each resource as they require. We let p_i denote the unit price of each resource

[3]Thus, if a client receives $R_{ij}x_j^\star$ amount of each resource i, her dominant share is $\max_i R_{ij}x_j^\star/b_i$, i.e., the bundle requirement μ_j divided by b.

i, so that each client j's per-job cost is $r_j = \sum_{i=1}^{m} p_i R_{ij}^{\gamma}$. Thus, the operator's total revenue is $\rho_r = \sum_{i=1}^{m} p_i \sum_{j=1}^{n} \left(R_{ij} x_j^{\star} \left(\sum_{l=1}^{m} p_l R_{lj} \right) \right)^{\gamma}$. The resource constraints, which require that the operator be able to meet all clients' resource requirements, are then $\sum_{j=1}^{n} R_{ij} x_j^{\star} \leq C_i$ for each resource i.

4.2 Comparing Pricing Plans

In this section, we compare the operator's revenue and/or fairness under bundled and resource pricing. We suppose that the operator chooses the prices so as to optimize a given objective function of clients' per-job costs and demands, e.g., the weighted sum $\nu\rho(\mathbf{p}) + F_{\beta}(\mathbf{p})$ of fairness and revenue. We use $f(\mathbf{r})$ to denote this function, and f_b^* and f_r^* to denote its maximum value under bundled and resource pricing respectively.

It is not obvious whether bundled or resource pricing will yield a higher value of f. One might expect that resource pricing would yield a higher value of f $\left(f_r^* \geq f_b^*\right)$, since with bundled pricing, the operator can choose only one price p while it chooses $m \geq 1$ resource prices p_i under resource pricing. However, with bundled pricing the operator can effectively force clients to purchase excess resources, which may lead to larger revenue.

To compare f_b^* and f_r^*, we first define a client's **dominant resource** as the resource k which maximizes R_{ij}/C_i over all resources i. We then find sufficient conditions under which bundled pricing does not perform better than resource pricing, i.e., $f_b^* \leq f_r^*$:

Lemma 1. *If the ratio of resources in each bundle is equal to the ratio of resource capacities, the maximum value of f under resource pricing may be at least as large as that under bundled pricing. Let μ^{γ_b} denote the (column) vector of clients' dominant shares raised to the bundled pricing volume discount γ_b, and let $\mathbf{R}^{\gamma_r}$ denote the resource matrix with (i,j) entries $R_{ij}^{\gamma_r}$. If $\mathbf{R}^{\gamma_r T} \left(\mathbf{R}^{\gamma_r} \mathbf{R}^{\gamma_r T} \right)^{-1} \mathbf{R}^{\gamma_r} \mu^{\gamma_b} = \mu^{\gamma_b}$ and $\left(\mathbf{R}^{\gamma_r} \mathbf{R}^{\gamma_r T} \right)^{-1} \mathbf{R}^{\gamma_r} \mu^{\gamma_b} \geq 0$, then $f_b^* \leq f_r^*$. These conditions are sufficient but not necessary.*

Proof. Let $p^{\star}$ be the optimal bundled price and consider resource prices $\mathbf{p} = p^{\star} \left(\mathbf{R}^{\gamma_r} \mathbf{R}^{\gamma_r T} \right)^{-1} \mathbf{R}^{\gamma_r} \mu^{\gamma_b}$ so that $\mathbf{R}^{\gamma_r T} \mathbf{p} = p^{\star} \mu^{\gamma_b}$, i.e., each client's per-job price r_j is the same under resource or bundled pricing. Each client's demand x_j is a function of r_j, so each client demands the same amount of jobs with and without bundling, satisfying the resource constraints. Since f can be written entirely in terms of clients' per-job costs r_j, the objective function value is also unchanged.

To show that this condition is not necessary, consider an example with three clients and two resources ($n = 3$, $m = 2$). Suppose that there is no volume discount ($\gamma = 1$), and that $f = \sum_{j=1}^{n} r_j x_j^{\star}\left(r_j\right)$, i.e., revenue. Let the demands $x_j^{\star}$ be determined by the isoelastic functions (1) with $c_j = 1$ for $j = 1,2,3$; $\alpha_1 = \alpha_3 = 0.5$, $\alpha_2 = 0.9$. We take the resource matrix and capacities to be $\mathbf{R} = \begin{bmatrix} 2 & 1 & 1 \\ 0.5 & 2 & 0.9 \end{bmatrix}$; $C_1 = C_2 = 10$, so that clients' dominant shares are $\mu_1 = \mu_2 = 0.2$ and $\mu_3 = 0.1$. Computing $\mathbf{R}\left(\mathbf{R}^T\mathbf{R}\right)^{-1}\mathbf{R}^T\mu = \begin{bmatrix} 0.191 & 0.189 & 0.130 \end{bmatrix}^T$, we see that $\mathbf{R}\left(\mathbf{R}^T\mathbf{R}\right)^{-1}\mathbf{R}^T\mu \neq \mu$. However, the optimal revenue under resource pricing is $f_r^* = 4.570$, which exceeds $f_b^* = 4.413$. □

For instance, Lemma 1's conditions are satisfied if each client has the same dominant resource and $\gamma_r = \gamma_b$. In this case μ^{γ} is simply a (transposed) scalar multiple of a row of $\mathbf{R}^{\gamma}$, and

$\mathbf{R}^{\gamma T}\left(\mathbf{R}^{\gamma}\mathbf{R}^{\gamma T}\right)^{-1}\mathbf{R}^{\gamma}\mu^{\gamma}=\mu^{\gamma}$. We can then find necessary and sufficient conditions under which bundled pricing is strictly worse than resource pricing $\left(f_b^* < f_r^*\right)$:

Proposition 3. *Suppose that all clients have the same dominant resource. Then the objective function value under resource pricing equals that under bundled pricing if the dominant resource yields the largest ratio of derivatives of the objective function and constraints. Let resource k be the dominant resource for all clients, let f be concave, and let the capacity constraints $\sum_{j=1}^{n} R_{ij}x_j^{\star}(\mathbf{r}) \leq C_i$ be convex, where r_j is client j's optimal price under bundled pricing. If, for each resource $i \neq k$,*

$$\left(\sum_{j=1}^{n}\frac{\partial f}{\partial r_j}R_{ij}^{\gamma}\right)\Big/\left(\sum_{j=1}^{n}R_{kj}\frac{\partial x_j^{\star}}{\partial r_j}R_{ij}^{\gamma}\right) < \left(\sum_{j=1}^{n}\frac{\partial f}{\partial r_j}R_{kj}^{\gamma}\right)\Big/\left(\sum_{j=1}^{n}R_{kj}\frac{\partial x_j^{\star}}{\partial r_j}R_{kj}^{\gamma}\right), \tag{2}$$

then $f_b^* = f_r^*$. Conversely, if (2) does not hold, then $f_b^* < f_r^*$.

Proof. Under bundled pricing, each client j pays a per-job price $r_j = \mu_j^{\gamma} p^{\star}$, where μ_j is client j's dominant share R_{kj}/C_k, γ is the volume discount, and $p^{\star}$ is chosen optimally. From the Karush-Kuhn-Tucker (KKT) necessary conditions for a local maximum, at the optimal bundled prices $\mathbf{r}$,

$$\frac{df}{dp} - \lambda\sum_{j=1}^{n}\mu_j\frac{dx_j^{\star}}{dp} = \sum_{j=1}^{n}\frac{\partial f}{\partial r_j}\mu_j^{\gamma} - \lambda\sum_{j=1}^{n}\mu_j\frac{dx_j^{\star}}{dr_j}\mu_j^{\gamma} = 0 \tag{3}$$

for a positive scalar λ that represents the multiplier of the bundled capacity constraint. We now consider resource pricing. From Lemma 1, we see that taking the resource prices $p_k = p^{\star}/C_k$, $p_i = 0$ for $i \neq k$ results in the same per-job prices r_j as under bundled pricing. Since we assume that the objective function f is concave and the capacity constraints convex, the KKT conditions are both necessary and sufficient for optimality in the resource pricing case. We next show that these are satisfied at the resource prices $p_i = 0$ for $i \neq k$, $p_k = p^{\star}/C_k$, i.e., that there exist multipliers $\lambda_i \geq 0$ for the capacity constraint of each resource i and $\nu_i \geq 0$ for the positivity constraint $p_i \geq 0$ that satisfy the KKT first-order optimality conditions.

We use λ/C_k as the multiplier for resource k's capacity constraint. Thus, the KKT condition for the price p_k is satisfied if the multiplier of each resource $i \neq k$'s capacity constraint $\sum_{j=1}^{m} R_{ij}x_j^{\star} \leq C_i$ equals zero. In fact, this must be the case, since these constraints are not tight (by assumption, resource k is the dominant resource of each client). Thus, it suffices to show that when we take λ/C_k to be the multiplier of resource k's capacity constraint and all other capacity constraint multipliers zero, we can find a positive scalar ν_i for each resource $i \neq k$ such that

$$\frac{\partial f}{\partial p_i} - \frac{\lambda}{C_k}\sum_{j=1}^{n}\frac{\partial x_j^{\star}}{\partial p_i} + \nu_i = \sum_{j=1}^{n}\frac{\partial f}{\partial r_j}R_{ij}^{\gamma} - \left(\frac{\lambda}{C_k}\right)\sum_{j=1}^{n}R_{kj}\frac{dx_j^{\star}}{dr_j}R_{ij}^{\gamma} + \nu_i = 0.$$

With λ defined by (3), we multiply by $\sum_{j=1}^{n} R_{kj}\frac{dx_j^{\star}}{dr_j}R_{kj}^{\gamma}$ to see that the constraint $\nu_i \geq 0$ is (2).

We prove the necessity of (2) by noting that from Lemma 1, taking $p_k = p^{\star}/C_k$ and $p_i = 0$ for all $i \neq k$ is a feasible solution to the resource pricing problem. Moreover, should this point be a local maximum, the multiplier ν_k of the inequality $p_k \geq 0$ must be zero, since $p_k > 0$. Thus, in order to satisfy the first-order KKT condition for p_k, we must have λ as in (3). But then if (2) is not satisfied, the KKT conditions are not satisfied and we do not have a local maximum. Thus, there must exist a set of feasible prices yielding higher revenue than these, and $f_r^* > f_b^*$. □

When clients do not share the same dominant resource, we derive sufficient conditions under which bundled pricing performs strictly worse than resource pricing:

Proposition 4. *Suppose that clients do not share the same dominant resource, i.e., for each resource i there exists a client j such that $R_{ij}/C_i < \mu_j$. Then if the conditions in Lemma 1 are satisfied and f does not have a local maximum at its optimal bundled price p_b, then $f_b^* < f_r^*$.*

Proof. Let p^* denote the optimal bundled price. From the proof of Lemma 1, there exists a price vector $\mathbf{p}$ such that for each client j, $p^* \mu_j^{\gamma_b} = \sum_{i=1}^m R_{ij}^{\gamma_r} p_i$, i.e., clients have the same per-job prices (and therefore the same demands $x_j^\star(\mathbf{p})$) under bundled and resource pricing. However, since f does not have a local maximum at p^*, the constraint under bundled pricing, $\sum_{j=1}^n \mu_j x_j^\star \leq 1$, is tight at p^*. But clients do not all share the same dominant resource, so none of the resource constraints $\sum_{j=1}^n R_{ij} x_j^\star \leq C_i$ are tight for any resource i. By continuity of f and $x_j^\star$, one can then find resource prices that increase the value of f without violating the resource constraints. □

We therefore show that under most, but not necessarily all, conditions, resource pricing performs at least as well as bundled pricing $\left(f_r^* \geq f_b^*\right)$.

5 Numerical Illustrations

In this section, we use a six-hour workload trace from a Google datacenter cluster to illustrate the effects on fairness and revenue of parameters not considered in Sections 3 and 4: the clients' resource requirements and volume discounts. In particular, we find that the heterogeneity of the clients' resource requirements strongly influences the tradeoff in fairness and revenue, and that increasing the volume discount allows an increase in revenue at the expense of decreased fairness.

5.1 Our Dataset

Our workload trace includes 9218 jobs divided into 176580 tasks; each task runs on a single machine within the cluster. The amount of memory (RAM) and fractional number of CPU cores taken up by each active task was recorded at five-minute intervals over 6 hours and scaled by a constant factor. For this reason, we omit units in our discussion here.

To simplify our simulations, we exclude jobs whose total usage of either CPU or memory lies more than one standard deviation away from the mean. We then use k-means clustering to group jobs into three different clusters or types by their CPU and memory usage and take the centroid points as the resource requirements: type 1 jobs require 0.4 CPUs and 2.7 units of memory, type 2 jobs require only 0.01 units of CPU and 0.02 units of memory, and type 3 jobs require 0.6 units of CPU and 0.5 units of memory. We associate a client type with each type of job and assume that these clients have isoelastic utility functions (1), with parameters $c_j = 1$ and $\alpha_j = 0.4, 0.7, 0.5$ for $j = 1, 2, 3$ respectively. Taking the capacity of each resource to be 6 units, the operator solves:

$$\max_{\mathbf{p}} \ \nu\rho(\mathbf{p}) + F_\beta(\mathbf{p}) \text{ s.t. } \{0.4x_1^\star + 0.01x_2^\star + 0.6x_3^\star,\ 2.7x_1^\star + 0.02x_2^\star + 0.5x_3^\star\} \leq 6. \tag{4}$$

5.2 Empirical Fairness-Revenue Tradeoffs

We first consider the effects of client heterogeneity on the operator's revenue and fairness and then consider the effects of volume discounts. Throughout our discussion, we take the fairness parameter β in (4) to be relatively large, at $\beta = 20$; this choice of β allows us to approximate **max-min fairness**, i.e., maximizing the minimum utility value. Larger values of β impose a stricter

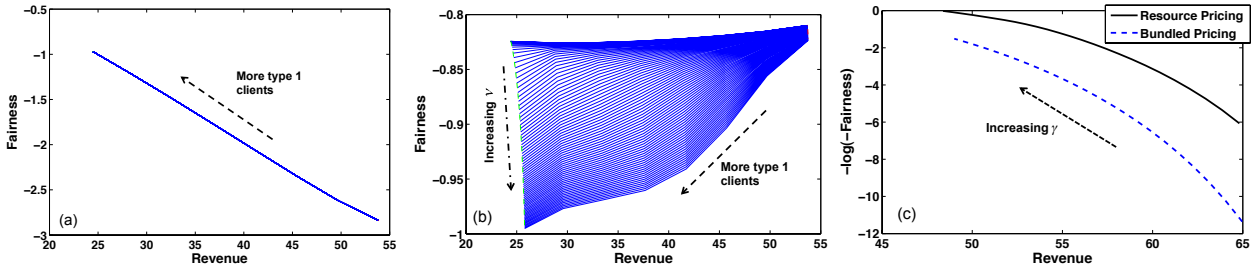

Figure 1: Achieved fairness and revenue for a range of revenue weights ν and (a) distribution of client types for bundled pricing, (b) distribution of client types for resource pricing, and (c) volume discounts γ. On each contour (individual contours not visible on the left and right), ν is fixed.

fairness requirement than lower ones, providing a fairness benchmark [7]. The units of fairness are therefore the same as the units of utility, i.e., dollars. For the purpose of exposition, we suppose that the datacenter can handle a total of 10 active clients in all numerical evaluations in this section.

We first examine the changes in fairness and revenue for different pricing plans when client types are less heterogeneous. We vary the fraction of type 1 and type 2 clients from 10 to 90% and fix the fraction of type 3 clients at 10%. Figure 1a shows the resulting fairness and revenue under bundled pricing; as expected from Prop. 1, for a given distribution of clients there is no fairness-revenue tradeoff. We see that there is a nearly linear relationship between fairness and revenue as the fraction of type 1 clients increases and that of type 2 clients decreases. This is because, as the fraction of type 1 clients increases, the revenue decreases and fairness increases: type 1 clients require more resources than type 2 clients, so the operator receives more revenue when its resources can be used to process type 2 clients' jobs. Thus, an operator's operating point is significantly affected by the characteristics of its clients.

Figure 1b shows that under resource pricing, the operator experiences a limited tradeoff between fairness and revenue (Prop. 2). Moreover, it can achieve both higher fairness and slightly higher revenue than under bundled pricing, as we would expect since Lemma 1's conditions are satisfied for these resource requirements. Increasing the weight ν in (4) yields a slight increase in the revenue and a limited decrease in fairness, though the fairness always remains higher than that for bundled pricing. As the fraction of type 1 clients increases, the revenue from both resource and bundled pricing decreases due to the discrepancy in resource requirements between type 1 and type 2 clients. Under bundled pricing (Figure 1a), the fairness increases with the fraction of type 1 clients: all type 1 clients receive the same (lower than type 2) utility values, so having more type 1 clients and fewer type 2 (high utility) clients increases the overall equitability, with only a few clients having utilities much higher than the average. Under resource pricing (Figure 1b), however, fairness decreases with the fraction of type 1 clients. Resource pricing allows the operator to allocate proportionally more resources to type 2 clients in order to increase its revenue. As the fraction of type 1 clients increases, this imbalance of resources is exacerbated, and type 2 clients' utility increases relative to type 1 clients' utility.

We conclude by examining the effect of the volume discount γ. For these simulations, we suppose that 80% of the clients are of type 2 and 10% each of types 1 and 3. Figure 1c shows the fairness-revenue tradeoffs for a range of volume discounts γ. Again, bundled pricing shows no tradeoff; moreover, there is almost no tradeoff even for resource pricing. Lemma 1's result that resource pricing yields a higher objective function value than bundled is demonstrated here

by the fact that revenue is comparable for both pricing plans, but resource pricing yields higher fairness. We also see that increasing γ, i.e., decreasing the volume discount, decreases revenue and increases fairness for both types of pricing. Smaller volume discounts dampen demand and thus revenue, especially from clients with large resource requirements, who experience proportionally larger changes in the per-job price with increased volume discounts. As demand from these clients decreases, proportionally more resources fall to clients with lower resource requirements, in this case type 2 clients. Since 80% of clients are of type 2, the overall fairness increases: most clients receive higher utility as the volume discount decreases (γ increases).

6 Conclusion and Future Work

In this work, we develop a framework for evaluating two cloud pricing strategies–bundled and resource pricing–in terms of their resulting fairness and revenue. We first characterize client demand for resources as a function of the prices offered under these different pricing plans. After showing some analytical bounds on the tradeoff between fairness and revenue, we compare achieved fairness and revenue under the two pricing plans. We finally use data taken from a Google cluster to numerically evaluate the impact of resource capacity and volume discounts on the operator's fairness-revenue tradeoff.

Future extensions of this work will consider an additional pricing scheme: differentiated pricing, in which the operator can choose a per-job price for each client independent of the client's resource requirements. We do not consider such a pricing plan here since bundled and resource pricing are more practically relevant; in practice clients are generally not charged different per-job prices. One could also extend our work to take into account job completion deadlines, which impose an additional constraint on the resources allocated at any given time. We also plan to consider tradeoffs between revenue, fairness, and operational efficiency, e.g., through examining the total amount of leftover resources.

References

[1] D. Bertsimas, V. F. Farias, and N. Trichakis. On the efficiency-fairness trade-off. *Mngmt. Sci.*, 2012.

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

[3] L. Du. Pricing and resource allocation in a cloud computing market. In *Proc. of IEEE/ACM CCGrid*, pages 817–822. IEEE, 2012.

[4] D. Harris. Google and Microsoft spent a combined $3.4b on infrastructure last quarter. GigaOm, 2013. http://tinyurl.com/k887stc.

[5] J. Hughes. Things to consider before choosing a data center. ExploreB2B, 2013. https://exploreb2b.com/articles/things-to-consider-before-choosing-a-data-center.

[6] IDC. Cloud computing in 2020. EMC, 2012. http://tinyurl.com/kma3lgc.

[7] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang. Multi-resource allocation: Fairness-efficiency tradeoffs in a unifying framework. In *Proc. of INFOCOM*, pages 1206–1214. IEEE, 2012.

[8] P. Koehler, D. Ma, A. Anandasivam, and C. Weinhardt. Customer heterogeneity and tariff biases in cloud computing. Proc. of the Intl Conf. on Information Systems (ICIS), 2010.

[9] 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.

[10] S. Wee. Debunking real-time pricing in cloud computing. In *Proc. of CCGrid*, pages 585–590, 2011.