Socially Optimal Pricing of Cloud Computing Resources

Ishai Menache
Microsoft Research New England
Cambridge, MA 02142
t-ismena@microsoft.com

Asuman Ozdaglar
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
asuman@mit.edu

Nahum Shimkin
Department of Electrical Engineering
Technion — Israel Institute of Technology
Haifa 32000, Israel
shimkin@ee.technion.ac.il

ABSTRACT

The cloud computing paradigm offers easily accessible computing resources of variable size and capabilities. We consider a cloud-computing facility that provides simultaneous service to a heterogeneous, time-varying population of users, each associated with a distinct job. Both the completion time, as well as the user's utility, may depend on the amount of computing resources applied to the job. In this paper, we focus on the objective of maximizing the long-term social surplus, which comprises of the aggregate utility of executed jobs minus load-dependent operating expenses. Our problem formulation relies on basic notions of welfare economics, augmented by relevant queueing aspects.

We first analyze the centralized setting, where an omniscient controller may regulate admission and resource allocation to each arriving job based on its individual type. Under appropriate convexity assumptions on the operating costs and individual utilities, we establish existence and uniqueness of the social optimum. We proceed to show that the social optimum may be induced by a *single* per-unit price, which charges a fixed amount per unit time and resource from all users.

Keywords

Cloud Computing, Pricing, Social Optimality

1. INTRODUCTION

Cloud computing is meant to offer on-demand network access to shared pools of configurable computing resources [14], such as virtual servers, applications and software services. This paradigm promises to deliver to the user the economics of scale of a large datacenter, fast and flexible provisioning of resources, and the freedom from long-term investments in equipment and related technology. The inner workings of the datacenter that supports the cloud operations are hidden from the user, who is presented with virtual servers, computing infrastructure, or software services. The idea of offering shared computing resources is of course not new, and has been extensively studied and implemented under different computing paradigms, including cluster, grid and utility computing [7]. Recently, however, the notion of cloud computing has gained prominence, spurred by numerous implementations by Amazon, Microsoft, Google, IBM and many others, both of public clouds (openly available over the Internet), or private cloud (intended for internal or restricted use). A recent survey of the promise and challenges of cloud computing can be for in [1], for example.

Shared computing facilities require effective mechanism for allocating available resource among users. This becomes a major challenge in view of the diversity of application types and user needs, which are at least partly hidden from the system manager. In public clouds, a major role is taken up by economic mechanisms, notably resource *pricing*, but also spanning more elaborate mechanisms such as bidding and auctions. Such economic mechanisms are naturally subject to revenue and profit considerations by commercial service provider. In this paper, however, our main focus is on the use of pricing as a means to maximize the *social welfare* associated with the cloud operations, which consists of the aggregate service utility obtained by the cloud users, less the infrastructure and operating costs of the service provider. Maximizing the social welfare (equivalently, the social surplus or social efficiency) is especially relevant for public clouds operated by a public organization or official agencies for the public benefit, as well as for private clouds set up by a commercial company or consortium of firms for internal use. Socially efficienct operation might as well be aligned with the interests of certain commercial public clouds, as it can help build the company's long-run reputation in this emerging market.

Social efficiency, and pricing as a means to achieve it, are basic notions in economic theory [11, 18]. The present paper leverages the standard theory, by considering explicitly the temporal dynamics of service. Our model considers a shared computing facility to which heterogeneous jobs (or applications) arrive sequentially. We assume that each job belongs to a distinct user, and henceforth use the terms *job and* user interchangeably. Each arriving user may acquire a certain amount of the computing resources, while all present jobs are served simultaneously, each on its allocated resources. The service quality experienced by the user naturally depends on the resources allocated to his job. Furthermore, depending on application, the job execution time may considerably scale down with the amount of resources applied to it. This is especially true for batch-type applications such as

Nov. 25, 2010; submitted to ValueTools 2011
.

scientific computing applications and business data analysis, that are computation and data intensive and can be efficiently parallelized. As argued in [1], such applications present notable opportunities for the utilization and further advancement of cloud computing. Both these factors, service quality and execution time, are essential parts of the user performance and utility model that we consider.

The pricing mechanism we consider is the simple usage-based pricing with linear tariffs, where each user pays a fixed amount per unit resource and unit time. Such pricing schemes are currently in use by several central cloud providers. The decision on how much resources to use and for what length of time is thus relegated to the user. Naturally, potential users may also decide to give up the offered service altogether, or *balk*. Such balking decisions effectively shape the arrival rate into the system, which together with the resource requirement of served users, determine the *demand* curve for the system resources. In the context of social welfare, pricing can be viewed as playing a dual role: First, regulating the overall system load to match available resources (or their operating costs), and second, inducing a distribution of available resources among users which is commensurate with their performance requirements and service utilities.

As mentioned, the suggested model emphasizes the sequential nature of user arrivals, and the dependence of their service times on the allocated resources. Accordingly, we develop the model in some detail, while clarifying the required assumptions. We then examine the social welfare under the assumption that system resources are sufficient to fully accommodate the socially optimal demand. Our main result is twofold: First, we establish the existence and uniqueness of the welfare maximizing solution (in terms of arrival rates and allocated resources to each job type). Second, we prove that there exists a unique price that induces this desired solution. We further elaborate on the relation of our model to the existing economic models. Some additional topics of interest that have been dealt with within this model include the issue of load constraints due to limited resources, consideration of profit, and iterative (tatonnement-like) price adjustment schemes that converge to the socially optimal price. These issues have been omitted here for lack of space and will appear in an extended version of this paper.

Let us briefly comment on related literature. Market-oriented mechanisms have been long been considered a means for resource allocation in shared computing systems, often focusing on system-oriented performance objectives such as average delay and throughput. A number of papers have considered the user-centric approach, where the objective is to maximize the aggregate service utility, using various market-based mechanisms and concepts, including commodity markets, bargaining, posted price models, contract based models, bid-based proportional resource sharing models, bartering, and various forms of auctions. Extensive surveys may be found in [19, 4]. The monograph [8] surveys related literature from a queueing theory perspective, while the monographs [13, 5] survey the use of pricing in telecommunications systems and communication networks.

A related body of work on bid-based resource allocation has emerged in the communication networks literature, in the context of capacity allocation and congestion control. Recent surveys may be found in [16, 20]. These bidding mechanisms may be viewed as adaptive, congestion-dependent pricing schemes, whereby the available resources are completely divided among the present users in proportion to their bids. An application of these concept to a shared (utility) computing environment has been considered in [21], where users are identifies as persistent flows of jobs, and bidding is employed to statically divide the computing resource among them. Our model here is basically different as users are associated with single jobs, which arrive sequentially and are allocated resources upon arrival.

The paper is organized as follows. Section 2 lays down the basic system and user model, and presents the required assumptions. Section 3 then considers the individual optimization problem faced by each arriving user, assuming a linear pricing schedule. In Section 4 we analyze the social optimization problem, and show that it admits a unique solution under our assumptions. The next Section 5 establishes our central result, namely that that social optimality is induced by fixed per-unit pricing, and characterizes the optimal price. The discussion in Section 6 puts the model and derived results in the context of the economic theory on social welfare. Finally, Section 7 presents some closing remarks and issues for further research.

2. SYSTEM AND USER MODEL

This section introduces our basic model. We start with the underlying service system, describe our assumptions regarding the service rates, and point out some steady-state relations. We further describe the users' utility functions. We finally specialize the general user model to the finite-class case, that will be treated in the rest of the paper.

2.1 The Service System

We consider a shared computing facility to which jobs (or software applications) are submitted sequentially by individual users. We associate each job with a distinct user, and employ the terms job and user interchangeably. Upon arrival, each job is allocated a certain amount of service resources, according to some resource allocation protocol, and promptly enters service which proceeds to completion. Alternatively, some potential arrivals may decide to balk (say, due to high pricing), in which case they leave the system without receiving any service.

We proceed to describe quantitatively the parameters of this model.

User types: Arriving users may differ in their service requirements and cost parameters. These are summarized by a *type* identifier, denoted i, which may be considered as a vector of real or discrete parameters. Let $\mathcal{I}$ denote the set of possible types.

Arrival rates: Potential users arrive according to some stochastic process, with specified rates for each type. with a given rate distribution. More precisely, let $\Lambda_0(di)$ denote a finite positive measure on $\mathcal{I}$. Then, for any (measurable) set $I \subset \mathcal{I}$, the arrival rate of potential users with types in I is $\Lambda_0(I)$. We impose no further requirement on the arrival processes except that the average number of arrivals converges in the long run to to specified averages[1] (almost surely), so that Little's law can be applied [17]. An arriving

[1]This allows to consider fairly general processes, that include, for example, non-stationarity due to time-of-day variation, Markov-modulated arrival processes, and dependence between the arrival processes of different types.

user may either choose to enter service, or else might balk with some probability (to be determined by our decision model). Thus, the *effective* arrival rates (sans the balking users), denoted $\Lambda(di)$, will generally be smaller than $\Lambda_0(di)$.

Resources: Let $z_i \geq 0$ denote the quantity of service resources allocated to a type-i user (or job) upon arrival. In our context, z_i may be thought of as the number of (virtual) computing units allocated to this job. We shall assume here that z_i is a continuous variable; this may indeed be the case in some systems, while for others it should be considered an approximation to a discrete variable with fine granularity. It is assumed that z_i is fixed throughout the job's execution period.

Service duration: Let τ_i denote the execution (service) time of some type-i job. As jobs are assumed to enter service upon arrival, this coincides with the job's sojourn time, namely the total time spent in the system. Evidently, the service time depends both on the service requirements of this job (as determined by the job type i), as well as the resources z allocated to that job. Thus, τ_i is a random variable with a z-dependent distribution, assumed to have finite mean and variance for any $z > 0$. We let $T_i(z) = E_{i,z}(\tau_i)$ denote the mean service time for type i jobs using z resources.

It will be naturally assumed that each $T_i(z)$ is a decreasing (or at least non-increasing) function of z. This property, along will some additional requirements on $T_i(z)$, will be formally stated in Subsection 2.6.

2.2 Steady State

We next consider the steady state load in the system for given arrival rates and resource allocations. Recall that type i users enter the system as a Poisson process with (effective) rate $\Lambda(di)$. Suppose that each type-i user is allocated a positive quantity z_i of resources. The service time is then a random variable with finite mean $T_i(z_i)$. As a result, the system may be viewed as a distribution over an independent collection (indexed by i) of $M/G/\infty$ queues, each of which is obviously stable. We shall assume that this system is in steady state. Using the sample-path version of Little's law [17], the long-term average number of jobs in service for a queue with arrival rate λ_i and mean service time $T_i(z_i)$ is given by $N_i = \lambda_i T_i(z_i)$. Therefore, the long-term average number of jobs is service (As a function of their type) is distributed according to

$$N(di) = \Lambda(di) T_i(z_i)\,.$$

As each type-i users occupies z_i resources, the long-term average of the total resource utilization is given by

$$Z = \int_i z_i N(di) = \int_i z_i T_i(z_i) \Lambda(di)\,. \tag{1}$$

We will refer to Z as the *load* of the system.

It may be noted that the description above presumes that the resource pool is unlimited, so that all arriving users are admitted to service. We will relate to the issue of limited resources later in Section **??**.

2.3 Pricing

Users will be charged some usage cost for the rendered service. We will focus here on fixed per-usage pricing, so that the monetary charge M_i for a job that occupied z_i resources for τ_i time units is

$$M_i = P z_i \tau_i\,,$$

where P is the per-unit price rate (in monetary units per unit resource and unit time). Therefore, the *expected* charge for a type-i job, given that it was allocated z_i resource, will be

$$E(M_i) = P z_i T_i(z_i)\,. \tag{2}$$

The latter expression will form part of the individual user's utility function. We note that the *expected* service time is used, which reflects the implicit assumption that an arriving user does not know beforehand the exact execution time of his job, but only its distribution.

2.4 Individual Utilities

An incoming user has two decisions to make upon arrival. One is whether to join the system or balk. If he decides to join, he further needs to determine the amount of resources z_i required for his job. These decisions are made individually by each user, with the goal of maximizing his own utility. We proceed to define the user utility function.

First, the utility of any balking user is taken to be zero. Note that this is merely a convenience, as any other baseline value can be used instead. As for users who join the system, the utility function of a type i user can be written as:

$$\begin{aligned} U_i(z_i) &= V_i(z_i) - E(M_i) \\ &= V_i(z_i) - P z_i T_i(z_i)\,, \end{aligned} \tag{3}$$

where V_i is the (expected) value that the user assigns to executing his job with resources z_i, and $E(m_i)$ is his expected charge as per (2).

The user value functions are assumed to satisfy to the following properties.

ASSUMPTION 1. *For each user type i,*

(i) $V_i(z)$ is continuously differentiable, and strictly concave increasing in $z \geq 0$, and bounded above.

(ii) $V_i(0) < 0$.

The concavity assumption is of course standard, implying that the marginal improvement due to additional resources is diminishing. Property (ii) simply guarantees that users will prefer balking to joining the system with $z = 0$ resources.

Discussion and Elaboration: It is important to note that the value V_i may reflect both the execution time of the job, as well as other elements related to the quality of service (QoS) experienced during the execution time. To make this specific, one may consider the separable form

$$V_i(z) = \breve{V}_i(z) - E_z(c_i(\tau_i))\,, \tag{4}$$

where the second component is the cost associated with the execution time, and the first represents the other quality measures. Here τ_i is the actual job execution time, $c_i(\cdot)$ is a delay cost function, and the expected value is taken with respect to the distribution of τ_i given $z_i = z$. We can now distinguish between two extreme cases:

• Batch jobs: Here a certain computation needs to be carried out, and the computation time scales with the allocated resources. Such jobs are common, for example, in scientific and business computing. Here $\breve{V}_i(z) = v_i$, independently of the allocated resources, and the delay sensitivity is the important term. Mild delay sensitivity may be captured by linear delay costs, namely, $c_i(\tau_i) = \gamma_i \tau_i$ for some $\gamma_i > 0$. Then $V_i(z) = \breve{V}_i(z) - \gamma_i T_i(z)$, where $T_i(z) = E_z(\tau_i)$ is the expected

service time of user i. Note that the assumed monotonicity and concavity properties of V_i are equivalent here to $T_i(z)$ being convex decreasing, which may be reasonably assumed (see below). Other applications may have more critical time constraints, which may be captured by a (convex increasing) function $c_i(\cdot)$ that becomes steep towards the required completion time.

• Fixed duration applications: In certain application classes, cloud resources may be secured for fixed periods of time. This may be the case, for example, in interactive applications that are intended for web customer service. In that case the delay term become irrelevant, and $V_i(z) = \breve{V}_i(z)$ captures the QoS offered to customers during that time period.

2.5 Operating Costs

Let C_{op} denote the operating cost of the computing facility per unit time. We assume that this cost depends on the system resource utilization, namely

$$C_{op} = C_0(Z)\,,$$

where Z denotes the *average* resource utilization, as specified in (1). We further assume the following.

ASSUMPTION 2. *$C_0(Z)$ is continuously differentiable and strictly convex increasing in $Z \geq 0$.*

Remarks

1. Large datacenters normally take advantage of the economy of scale offered by statistical multiplexing and resource virtualization, so that the overall load on the system is smaller than the sum of individual resource requirements. The cost function C_0 is assumed to take account of this effect.

2. In addition to the running costs of operation, the cost term may take into account also also the required investment in infrastructure, computed for a certain period ahead. Whether this is included depends of course on the time scale considered, and whether investment in infrastructure is considered as part of the model.

3. It may be argued that the operating costs are better described in terms of the instantaneous resource utilization, in the form $E(C_0(\tilde{Z}))$, where $\tilde{Z}$ is distributed as the steady-state resource utilization. While this may be the case to some extent, we note the following:

   a. System operation and related expenses are often dictated by the average utilization. For example, a decision as to whether to put into operation a (possibly high-cost) standby facility will typically be made based on average expected utilization rather than short-terms spikes. This is better represented by $C_0(Z)$.

   b. The additional costs that are associated with the instantaneous operation of each unit may well be approximated by a linear cost component that fluctuates around the average, so that $C_0(\tilde{Z}) \simeq C_0(Z) + c_1(Z)(\tilde{Z} - Z)$. Taking the expected value and noting that $E(\tilde{Z}) = Z$, we obtain $E(C_0(\tilde{Z})) = C_0(Z)$.

2.6 Service Time

As mentioned, the service time, or job execution time, generally depends on the resources z allocated to it. Recall that $T_i(z)$ denotes the mean service time function for type-i jobs. We discuss here some specific forms for these functions, and then state our general assumptions.

With the exception of fixed-duration applications, we reasonably expect the mean service time to be strictly decreasing in z. A common assumption in the processor-sharing queueing literature is that of proportional speedup, namely $T(z) = D/z$. This basic model is arguably overly optimistic regarding the benefits of scale in parallel computation, as it ignores factors such as setup time and parallelization overhead that should impede further reduction in T beyond a certain point. A slightly modified model that can accommodate such effects is given by

$$T(z) = a + \frac{D}{z}\,. \tag{5}$$

Here $a > 0$ presents the non-scalable part of the job execution. Obviously, now $T(\infty) = a > 0$. This model has the same form as Amdahl's law, which is often used to model possible speedup in parallel computing (e.g., [9]).

Our general requirements on $T_i(z)$ are given below, and involve the user value functions $V_i(z)$ as well. We will subsequently state more specific conditions on T_i that imply this assumption.

ASSUMPTION 3. *For each user type i, $T_i(z)$ satisfies the following properties:*

(i) $T_i(z)$ is a continuously differentiable and (weakly) decreasing function of $z \geq 0$, with $\lim_{z\to\infty} T_i(z) > 0$.

(ii) The ratio $\frac{V_i'(z)}{(zT_i(z))'}$ is strictly decreasing in z. Here V_i is the user utility function, and the primes denote differentiation with respect to z.

Observe that these conditions are satisfied for the model in (5). In that case, $zT_i(z) = az + D$ and $(zT_i(z))' = a$, a positive constant, so that (ii) is equivalent to concavity of $V_i(z)$ (which is indeed included in our assumptions). The case of a fixed execution time is of course a special case with $D = 0$. More generally, property (ii) holds whenever $zT_i(z)$ is a convex increasing function of z, as this implies that the denominator is non-decreasing and positive.

Convexity of $zT_i(z)$ is however not a necessary condition for Assumption 3 to hold. As an important example, property (ii) above can be established for certain types of delay-sensitive utility functions, provided that the expected service time satisfies some reasonable additional conditions on the service rates. We summarize this observation in the following lemma.

LEMMA 1. *Let $\mu_i(z) = 1/T_i(z)$ denote the* service rate *function. Suppose that*

(i) $\mu_i(z)$ is a differentiable, strictly concave and strictly increasing function of $z \geq 0$, with $\mu_i(0) = 0$ and $\mu_i(\infty) < \infty$.

(ii) $V_i(z) = v_i - c_i(T_i(z))$, where c_i is an increasing and convex function of z.

Then Assumption 3 is satisfied.

PROOF. Item (i) of the assumption is obvious by the stated properties of μ_i. Let $f(z) \stackrel{\triangle}{=} \frac{\bar{V}'(z)}{(zT(z))'}$ (where we omit the index i). Substituting $V(z) = v - c(T(z))$ and $T(z) = \mu(z)^{-1}$,

we obtain

$$f(z) = \frac{c'(\mu(z)^{-1})\mu'(z)}{\mu(z) - z\mu'(z)} \triangleq \frac{N(z)}{D(z)}\,.$$

Item (ii) now follows by showing that $N(z)$ is positive decreasing and $D(z)$ is positive increasing (with both monotonicity properties being strict). The first claim follows since $\mu(z)^{-1}$ is strictly decreasing, $c'(\cdot)$ is positive increasing, and $\mu'(z)$ is positive strictly decreasing (by the assumed concavity of μ). Monotonicity of $D(z)$ follows by observing that $D'(z) = -z\mu''(z) > 0$ for $z > 0$ (where one-sided derivatives may be used if necessary), and positivity now follows since $D(0) = 0$. $\square$

To illustrate, $T(z) = a + \frac{D}{\sqrt{z}}$ does not satisfy convexity of $zT(z)$, but condition (i) of the last lemma is easily seen to hold.

2.7 The Finite Class Model

The model as described above allows a continuum of user types, with different performance and utility characteristics for each type. While possible to continue the analysis at this level of generality, we find it useful to consider here a finite dimensional model, that is amenable to explicit computations and avoids technicalities associated with measurability issues. The first such model that comes to mind is restricted to a *finite* number of user types, each with a positive mass of arrivals. However, such a model suffers from a couple of shortcoming, both related to the admission decisions of the users.

1. Discontinuous demand: Consider the variation in the arrival rate as the price is increased. As all users of a given type share identical parameters, they will all change their admission decisions (from join to balk) at the same price level. This will lead to jumps in the demand, in response to some small changes in price. Such discontinuities can hardly be expected in practice.

2. Mixed decisions: Due to the above-mentioned discontinuities, equilibrium conditions will generally require users of one or more types to choose probabilistically between join or balk. As such users are necessarily neutral with respect to these choices, the precise mechanism through which a given user comes to choose between them with a given probability remains exogenous to the model.

To nullify these shortcomings, we will consider a *finite-class* model that goes beyond the simple finite-type case. Here users are grouped into a finite set of *user-classes*, each sharing the same characteristics except for a continuously-distributed bias is their service utility. As we shall see, this addition will indeed induce smooth demand variation, and avoid randomized decisions. The main characteristics of this model are borrowed from [12], where a similar utility model was used in a queueing context.

Let the set $\mathcal{I}$ of user types be divides into a finite set of classes, denoted $\mathcal{S} = \{1, \ldots, S\}$, with elements $s \in \mathcal{S}$. We use the notation $i \in s$ to indicate that a type i belongs to class s. All jobs of a given class s have similar service time characteristics, namely

$$T_i(z) \equiv T_s(z)\,, \quad \text{for all } i \in s\,.$$

Furthermore, the service value functions $V_i(z)$ are taken to have the additive form[2]

$$V_i(z) = v_i + V_s(z)\,, \quad i \in s\,. \tag{6}$$

Thus, the dependence on the resource z is the same for all users of a given class. To that, a type-dependent bias v_i is added which creates intra-class variation. We refer to v_i as the user *taste* parameter.

Combining the above with (3) and (2), the utility function of a served user is

$$U_i(z, P) = v_i + V_s(z) - PzT_i(z)\,. \tag{7}$$

The user type i may now be identified with the pair (s, v), namely the users' class and his taste parameter. Recall that the potential arrival rates are specified through a positive measure $\Lambda_0(di)$ on the set of types $\mathcal{I}$. With $i = (s, v)$, we may express $\Lambda_0(di)$ as $\Lambda_0(s, dv)$; here $\Lambda_0(s, \cdot)$ is the distribution of the taste v for class s users. Let $\lambda_s^{\max} = \Lambda_0(s, \mathbb{R})$ denote the total arrival rate of potential users of class s. Some further requirements regarding these distributions will be specified in Assumption 5 below.

We observe that Assumptions 1 and 3 regarding V_i and T_i are in effect, and these imply similar properties for the class quantities T_s and V_s. We summarize these properties below.

ASSUMPTION 4. *For each user class s,*

(i) *$V_s(z)$ is continuously differentiable, strictly concave increasing for $z > 0$, and bounded above.*

(ii) *$V_s(0+) = -\infty$.*

(iii) *$T_s(z)$ is a continuously differentiable and decreasing function of $z \geq 0$, with $\lim_{z\to\infty} T_s(z) > 0$.*

(iv) *The ratio $\frac{V_s'(z)}{(zT_s(z))'}$ is strictly decreasing in z.*

Property (ii) ensures that Assumption 1(ii) is satisfied for any taste parameter v_i.

2.8 Aggregate Utility

Given the utility function in (7), we obtain the following *demand function*

$$\lambda_s(z, P) = \lambda_s^{\max} Prob\{v + V_s(z) - PzT_i(z) \geq 0\}\,,$$

where the probability is taken over v according to its class distribution $\Lambda_s(s, dv)$. Thus, $\lambda_s(z, P)$ is the arrival rate of users whose utility is non-negative when allocated z resources at price P. More important for our purpose, however, will be the aggregate utility obtained at a given arrival rate. Suppose that, out of the potential arrivals of class s, only those users with higher taste parameter v are admitted up to rate $\lambda_s \in [0, \lambda_s^{\max}]$, and each of those is allocated $z_s > 0$ resources. The aggregate value of service for these admitted users (per unit time) will be

$$\mathcal{V}_s(\lambda_s, z_s) = \bar{V}_s(\lambda_s) + \lambda_s V_s(z_s)\,, \tag{8}$$

[2]To avoid notational clutter, here and in the following we distinguish between some type-specific and class-specific quantities (such as V_i as V_s) through their index only.

where

$$\bar{V}_s(\lambda_s) = \sup_{\{e_v \in [0,1]\}} \left\{ \int v e_v \Lambda_0(s, dv) \ : \ \int e_v \Lambda_0(s, dv) = \lambda_s \right\} . \tag{9}$$

Thus, $\bar{V}_s(\lambda_s)$ is the sum over the higher-percentile tastes of users of class s, up to rate λ_s. We refer to $\bar{V}_s$ as the (taste) *aggregate utility*. A more explicit expression is obtained as follows. For each λ_s, let v_0 and $p \in [0, 1)$ be so that

$$\lambda_s = \int_{v > v_0} \Lambda_0(s, dv) + p\Lambda_0(s, v_0) \tag{10}$$

(note that $p \neq 0$ is required only if $\Lambda_0(s, \cdot)$ has a point mass at v_0). Then

$$\bar{V}_s(\lambda_s) = \int_{v > v_0} v\Lambda_0(s, dv) + pv_0\Lambda_0(s, v_0) \, . \tag{11}$$

We next discuss some properties of the aggregate utility functions $\bar{V}_s(\lambda_s)$. It is easily verified that $\bar{V}_s(0) = 0$, and $\bar{V}_s$ is (weakly) concave. With some additional assumptions it satisfies stricter properties (cf. [12]).

LEMMA 2. *For each class s, consider the function $\bar{V}_s(\lambda)$ defined in (9). Suppose that the measure $\Lambda_0(s, dv)$ over v is absolutely continuous (relative to the Lesbegue measure), with a density function $g_s(v)$. Then*

(i) $\bar{V}_s(\lambda)$ is strictly concave in $\lambda \in [0, \lambda_s^{\max}]$.

(ii) Suppose $g_s(v) > 0$ in some neighborhood of v_0. Then $\bar{V}_s(\lambda)$ is continuously differentiable around $\lambda_0 = \int_{v \geq v_0} g_s(v)dv$, with the derivative $\bar{V}_s' \stackrel{\triangle}{=} \frac{d\bar{V}_s}{d\lambda}$ given by $\bar{V}_s'(\lambda_0) = v_0$.

(iii) If $g_s(v) > 0$ for all $-\infty < v < \infty$, then $\bar{V}_s(\lambda)$ is continuously differentiable for all $\lambda \in [0, \lambda_s^{\max}]$, and $\lim_{\lambda \to 0} \bar{V}_s'(\lambda) = \infty$, $\lim_{\lambda \to \lambda_s^{\max}} \bar{V}_s'(\lambda) = -\infty$.

PROOF. (i) Given the existence of density, (10)-(11) takes the form

$$\lambda = \int_{v \geq v_0} g_s(v)dv \, , \tag{12}$$

$$\bar{V}_s(\lambda) = \int_{v \geq v_0} v g_s(v)dv \, . \tag{13}$$

In fact, for every $\lambda \in (0, \lambda_s^{\max})$ there exists a some $v_0 = v_0(\lambda)$ that satisfies (12). This follows by noting that the right-hand side is continuous in v_0 and (weakly) decreasing from $\lambda_s^{\max}$ to 0 as v is increased from $-\infty$ to ∞. Now, using (13), we have that for every λ and $\epsilon > 0$,

$$\begin{aligned}
\bar{V}_s(\lambda + \epsilon) - \bar{V}_s(\lambda) &= \int_{v_0(\lambda+\epsilon)}^{v_0(\lambda)} v g_s(v)dv \\
&> v_0(\lambda) \int_{v_0(\lambda+\epsilon)}^{v_0(\lambda)} g_s(v)dv \\
&= v_0(\lambda)\epsilon = v_0(\lambda) \int_{v_0(\lambda)}^{v_0(\lambda-\epsilon)} g_s(v)dv \\
&> \bar{V}_s(\lambda) - \bar{V}_s(\lambda - \epsilon) \, ,
\end{aligned}$$

which implies strict concavity.

(ii) From $g_s(v) > 0$, it follows that $v_0 = v_0(\lambda)$ that satisfied (12) is continuous and strictly decreasing in λ around λ_0. Observe now from (12)-(13) that $\frac{d\lambda}{dv_0} = -g_s(v_0)$ and $\frac{d\bar{V}_s}{dv_0} = -v_0 g_s(v_0)$, so that $\frac{d\bar{V}_s}{d\lambda} = v_0(\lambda)$ around λ_0. Since $v_0(\lambda)$ is continuous there, then so is the latter derivative.

(iii) The first part follows from (*ii*). The limits follow from $V_s'(\lambda) = v_0(\lambda)$, as (12) implies that $v_0(\lambda) \to \infty$ as $\lambda \to 0$, and that $v_0(\lambda) \to -\infty$ as $\lambda \to \lambda_s^{\max}$. □

We will henceforth adopt the following assumption.

ASSUMPTION 5. *The conditions of Lemma 2 are satisfied. That is, for each class s, the taste distribution $\Lambda_0(s, dv)$ admits a finite density g_s, with $g_s(v) > 0$ for all[3] $-\infty < v < \infty$.*

We finally note that, under the same conditions that lead to (8), the average load of equation (1) can be expressed as

$$Z = \sum_s \lambda_s z_s T_s(z_s) \, . \tag{14}$$

3. INDIVIDUALLY OPTIMAL DECISIONS

We proceed to examine the optimization problem faced by an individual user. Given the advertised per-unit price P, an arriving user needs to decide whether to execute his job at the considered facility, and if so, the amount of resources to acquire for that purpose.

Consider a user i of class s and taste v_i. Recall that this user's utility function is given by (7) if he joins service, and set to 0 if he chooses to balk. In this section we consider a fixed price $P > 0$. We therefore omit P from our notation and write $U_i(z)$ for $U_i(z, P)$, etc. The maximal utility for this user will be

$$U_i^{\max} = \max\{0, \max_{z \geq 0} U_i(z)\} \, , \tag{15}$$

$$= \max\{0, v_i + U_s^*\} \tag{16}$$

where

$$U_s^* = \max_{z \geq 0}\{V_s(z) - PzT_s(z)\} \, . \tag{17}$$

As we show below, the optimization problem in (17) admits a unique maximum. Recall that a scalar function the real line is *strictly quasiconcave* if it is strictly increasing up to a certain point, and strictly decreasing thereafter.

PROPOSITION 3. *The utility function $U_s(z) \stackrel{\triangle}{=} V_s(z) - PzT_s(z)$ is strictly quasiconcave, and admits a unique maximizer $z_s > 0$, which satisfies the following first-order conditions*

$$V_s'(z_s) - P\left(z_s T_s(z_s)\right)' = 0 \, . \tag{18}$$

PROOF. We first observe that $U_s(z)$ has at most one stationary point z where $U_s'(z) = 0$. Indeed the latter is equivalent to

$$\frac{V_s'(z)}{(zT_s(z))'} = P.$$

But $U_s(0+) = -\infty$ by Assumption 4(ii), and $U_s(z)$ is eventually decreasing since $\lim_{z \to \infty} zT_s(z) = \infty$ by Assumption 3(i), while V_s is bounded from above. Consequently, there must exist a maximum point at some finite $z > 0$, which clearly must satisfy $U_s'(z) = 0$, namely (18). Since there is no other stationary point, the assertion follows. □

[3]The infinite support of g_s is a modeling convenience, as it ensures that at any price level there will be some users which choose to enter service, and some others who choose to balk. In reality, we need this property to hold only for prices in a reasonable range.

We may summarize user $i's$ decision process as follows. First, he computes the optimal resource allocation z_s and maximal utility $U_s^* = U_s(z_s)$ by solving (17). Next, if $v_i + U_s(z_s) < 0$ he balks, if > 0 he enters service using z_s resources, and in case of equality he is neutral between these two options. For concreteness we shall choose the enter option in this case.[4]

We can now obtain the effective arrival rate λ_s of each user class. As noted, users of class s who join service are those with tastes $v_i \geq -U_s(z_s)$, where $U_s(z) = V_s(z) - P^* z T_s(z)$. Then λ_s is given by (10) with $v_s = -U_s(z_s)$, and, observing Lemma 2(ii), $\bar{V}_s'(\lambda_s) = v_s$, or

$$\bar{V}_s'(\lambda_s) + V_s(z_s) - P^* z_s T_s(z_s) = 0\,. \tag{19}$$

Since $\bar{V}_s'$ is a strictly increasing function, this equation uniquely determines λ_s.

We finally establish some plausible monotonicity properties, that will be be needed later on. Let $z_s(P)$ and $\lambda_s(P)$ denote the individually optimal resource allocation and effective arrival rate under price P.

LEMMA 4. *$z_s(P)$, $z_s(P)T_s(z_s(P))$ and $\lambda_s(P)$ are all continuous and strictly decreasing functions of P. Consequently, so is $Z(P) = \sum_s \lambda_s z_s T_s(z_s)$.*

PROOF. Fixing s, we remove the class index from V_s, $z_s(P)$, etc. in the remainder of this proof. Considering $z(P)$, we write (26) as

$$P = \frac{V'(z)}{(zT(z))'}\,. \tag{20}$$

But the right-hand side is continuous and strictly decreases in z (by Assumption 4), so that $P = P(z)$ is strictly decreasing in z. This implies that the inverse function $z_s = z_s(P)$ is a well-defined continuous function which is strictly decreasing in P.

Turning to $z(P)T(z(P))$, note that $(zT(z))' > 0$ at any solution $z = z(P)$ of (20), since $V' > 0$ (Assumption 4) and $P > 0$. Therefore $h(z) \stackrel{\triangle}{=} zT(z)$ is strictly increasing in z at these points. But we have just shown that $z(P)$ is strictly decreasing in P, and therefore so is $h(z(P))$. Continuity of the latter follows from that of $z(P)$ and $T(z)$.

Finally, consider $\lambda(P)$. By (25),

$$\bar{V}'(\lambda) = -V(z) + P\,T(z)z \stackrel{\triangle}{=} k(z)\,,$$

so that $\bar{V}'(\lambda(P)) = k(z(P))$. Recall that $\bar{V}'(\lambda)$ is continuous and strictly decreasing in λ by Lemma 2. Therefore, the required properties of $\lambda(P)$ would follow by showing that $k(z(P))$ is continuous and strictly increasing in P. But

$$\frac{dk(z(P))}{dP} = \frac{\partial k}{\partial z}\frac{dz}{dP} + \frac{\partial k}{\partial P} = \frac{\partial k}{\partial P} = T(z(P))z(P) > 0$$

where we have used (26) to conclude that $\frac{\partial k}{\partial z} = 0$ (which is of course related to the Envelope Theorem [18]). This concludes the proof. □

The last lemma shows that, sensibly, as the price P is increased, users will acquire less resources z_s, and their arrival rates λ_s will decrease. Further, The multiple $z_s T_s(z_s)$ that represents the total resource usage over time by each user is decreasing as well (even though the execution time $T_s(z_s)$ increases under our assumptions), as does the average system load Z.

[4]As the set of neutral users will always have zero measure under our type assumptions, this choice does not affect our results.

4. THE OPTIMAL SOCIAL WELFARE

The social welfare, or social utility, is defined as the sum of utilities of all individual entities that are considered part of the society. In our model these are the individual users together with the service provider. We consider here the socially optimal assignment of arrivals and resource allocation, which is intended to maximize the social welfare. We allow this assignment to be managed by an omniscient central controller, that has full knowledge of the the system parameters as well as individual customer types and preferences. This is not a realistic scenario of course, and is used only to identify the social optimum. Later we will show that this optimum can be achieved by appropriate pricing.

We proceed to present the social welfare function that is to be maximized, and characterize its optimal solution. We start by presenting a general expression for the social welfare, which we then specialize to the finite-class model. The last subsection establishes existence and uniqueness of the optimal solution.

4.1 The Social Welfare

The controller's decisions may be expressed in terms of the following variables:

1. $e_i \in [0,1]$, where $e_i = 1$ means that users of type i are admitted to service, $e_i = 0$ means rejection, and $e_i \in (0,1)$ means randomization between these two options.[5]

2. $z_i \geq 0$, the amount of resources to assign to users of type i (which is relevant only if $e_i > 0$).

Recall that the potential arrival rate distribution is specified by $\Lambda_0(di)$. The social welfare is now given by the sum of user utilities minus the system operating expenses:

$$W_{soc} = \int_i V_i(z_i) e_i \Lambda_0(di) - C_0(Z) \tag{21}$$

where

$$Z = \int_i z_i T_i(z_i) e_i \Lambda_0(di)\,.$$

Note that this expresses the steady-state expected social surplus per unit time, or equivalently its long-term average. Our goal is to maximize W_{soc} over all (measurable) selections of decision variables. Let W_{soc}^* denote this maximal value.

4.2 Aggregate Utility Form

Specializing to the finite-class (but infinite-taste) model, we proceed to formulate the above optimization problem as a finite dimensional mathematical program. Suppose within each class s only users with higher tastes are admitted, up

[5]We assume that all users of the same type are subject the the same control decisions. This can be argued to be optimal; however we will not bother with that here since under a continuum of types assumption the chances of obtaining two users of the same type are null. Furthermore, randomized decisions will not be required in this case.

to rate λ_s, and each of these is allocated resources z_s. Then, observing (8) and (14), the social welfare rate is given by

$$W(\boldsymbol{\lambda}, \mathbf{z}) = \sum_s \ \bar{V}_s(\lambda_s) + \lambda_s V_s(z_s) \ - C_0 \ \sum_s \lambda_s T_s(z_s) z_s \ , \tag{22}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)$, $\mathbf{z} = (z_1, \ldots, z_s)$. Recalling the definition of $\bar{V}_s$, this can be interpreted as the social welfare obtained when admitting class-s users with higher tastes up to rate λ_s, and allocating z_s resources to each. Consider the following optimization problem:

$$\begin{aligned} &\text{maximize } W(\boldsymbol{\lambda}, \mathbf{z}) \\ &\text{subject to} \quad \lambda_s \in [0, \lambda_s^{\max}], \quad s \in \mathcal{S}, \\ &\qquad\qquad\ \ z_s \geq 0, \qquad\qquad s \in \mathcal{S}. \end{aligned} \tag{23}$$

PROPOSITION 5. *The maximal value W^*_{soc} of the social welfare* (21) *coincides with the optimal value of the program* (23).

PROOF. We first argue that all users of a given class can be allocated identical resources. This is most easily demonstrated using the first variation of (21). Consider the maximization of (21) over $(z_i)_{i \in s}$ for a single class s, with all other decision variables fixed. Substituting an ϵ-variation $z_i^\epsilon = z_i + \epsilon \tilde{z}_i$, we obtain after some calculation

$$W_{soc}(\mathbf{z}^\epsilon) = W_{soc}(\mathbf{z}) + \epsilon \int_{i \in s} [V_s'(z_i) - C_0'(Z)(z_i T(z_i))'] \tilde{z}_i e_i \Lambda(di) + o(\epsilon) .$$

Note that we substituted $V_i' = V_s'$, which follows from (6). In any maximum the variation term must be non-positive. Now, for $z_i > 0$, $\tilde{z}_i$ can have arbitrary sign, so that

$$V_s'(z_i) - C_0'(Z)(z_i T(z_i))' = 0, \quad z_i > 0$$

must hold for $e_i \Lambda(di)$-almost every $i \in s$. If $z_i = 0$ then $\tilde{z}_i$ is non-negative (assuming $\epsilon > 0$), and we similarly obtain

$$V_s'(z_i) - C_0'(Z)(z_i T(z_i))' \leq 0, \quad z_i = 0 .$$

Noting that $C_0' > 0$ by Assumption 2, if follows as in Proposition 3 that the last two equations have a unique solution $z_i \equiv z_s$, which is valid for all $i \in s$ with $e_i > 0$ (i.e., which are admitted to service). We can therefore restrict attention to $z_i \equiv z_s$ for all $i \in s$. Substituting in (21) gives, after noting (6) and (14),

$$\begin{aligned} W_{soc} &= \int_{i=(s,v)} [v + V_s(z_s)] e_i \Lambda_0(di) C_0 \ \sum_s \lambda_s z_s T_s(z_s) \Big) \\ &= \sum_s \ \int_v v e_{(s,v)} \Lambda_0(s, dv) + \lambda_s V_s(z_s) \\ &\quad - C_0 \ \sum_s \lambda_s z_s T_s(z_s) \Big) \end{aligned}$$

where $\lambda_s = \int_{i \in s} e_i \Lambda_0(di) = \int_v e_{(s,v)} \Lambda_0(s, dv)$ is the *effective* arrival rate of class i.

Consider now the maximization of the last expression for W_{soc} over $\{e_i \equiv e_{(s,v)}\}$. For given λ_s, the only term that is sensitive to the choice of $e_{(s,v)}$ is the first one, and its maximum is clearly obtained by $\bar{V}(\lambda_s)$ in (9). With this substitution, W_{soc} reduces to the expression in (22). □

4.3 Existence and Uniqueness

We proceed to show existence and uniqueness of the solution to the social optimization problem (22)-(23), and characterize this solution in terms of the first-order optimality conditions. We note that this program is not a concave one, even under the convexity properties imposed in our model assumptions. In fact, it is readily seen that this program is a convex one in $\boldsymbol{\lambda}$ alone (with $\mathbf{z}$ held fixed), and the the optimization problem over $\mathbf{z}$ can be transformed into a convex one (as discussed in Section 6). However, the problem is essentially *not* jointly convex in $\boldsymbol{\lambda}$ and $\mathbf{z}$, due to the multiplicative terms $\lambda_s V_s(z_s)$ and $\lambda_s T_s(z_s) z_s$. Hence, we resort to problem-specific analysis that relies on monotonicity arguments.

We start with the following characterization of the (possibly local) maxima of our optimization problem.

LEMMA 6. *Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)$, $\mathbf{z} = (z_1, \ldots, z_S)$ be a local maximum point of* (23)*, and define*

$$\eta \equiv \eta(\boldsymbol{\lambda}, \mathbf{z}) = C_0' \ \sum_s \lambda_s z_s T_s(z_s) \ . \tag{24}$$

Then $\lambda_s \in (0, \lambda_s^{\max})$, $z_s > 0$, and

$$\bar{V}_s'(\lambda_s) + V_s(z_s) - \eta \, T_s(z_s) z_s = 0 , \quad s \in \mathcal{S} , \tag{25}$$

$$V_s'(z_s) - \eta \ z_s T_s(z_s) \ ' = 0, \quad s \in \mathcal{S} . \tag{26}$$

PROOF. If follows from Lemma 2(iii) that λ_s is internal, namely $\lambda_s \notin \{0, \lambda_s^{\max}\}$. Differentiating (22) with respect to λ_s yields (25). Now, $\lambda_s > 0$ implies that $z_s > 0$ upon noting Assumption 4(ii). Thus, the maximizing z_s is interior. Equating the derivative of (22) with respect to z_s to zero and cancelling λ_s yields (26). □

We next shown that equations (24)–(26) admit a unique solution. This will follow by showing that equations (25)–(26) imply that $\{\lambda_s, z_s\}$ are decreasing in η, while the right-hand side of (24) is increasing in these variables. The required monotonicity properties of $\{\lambda_s, z_s\}$ are summarized in the next lemma.

LEMMA 7. *Consider equations* (25)–(26)*, with fixed $\eta > 0$.*

(i) For any $\eta > 0$, there exists a unique solution $\{\lambda_s, z_s\}$ to equations (25)–(26)*. Denote this solution by $\{\lambda_s(\eta), z_s(\eta)\}$.*

(ii) The functions $z_s(\eta)$, $z_s(\eta) T_s(z_s(\eta))$ and $\lambda_s(\eta)$ are all continuous and strictly decreasing in η.

PROOF. (i) Fix $\eta > 0$ and s. Existence and uniqueness of a solution z_s to equation (26) follows as in Proposition 3. Existence and uniqueness of a corresponding solution λ_s to (25) now follows by the properties of $\bar{V}_s$ in Lemma 2, items (i) and (iii).

(ii) The proof is identical to that of Lemma 4. □

LEMMA 8. *There exists a unique solution $\{\lambda_s^*, z_s^*\}$ to the system of equations* (24)–(26).

PROOF. Consider the right-hand side of equation (24) as a function of $\eta > 0$, with $z_s = z_s(\eta)$ and $\lambda_s = \lambda_s(\eta)$ as specified in Lemma 7. By the results of that lemma, the argument of C_0' is strictly decreasing in η, and since C_0' is a strictly increasing function (by the assumed convexity of $C_0(Z)$) it follows that $C_0'(\sum_s \lambda_s z_s T_s(z_s))$ is strictly decreasing in η. Since it is also positive, it follow that (24) has a unique solution η^*, with corresponding $z_s^* = z_s(\eta^*)$ and $\lambda_s^* = \lambda_s(\eta^*)$. □

We finally need to show that the maximum of (23) is obtained in a compact set, namely not for $z_s \to \infty$.

LEMMA 9. *The global maximum of* (23) *is attained at a finite point.*

PROOF. We show that $W(\boldsymbol{\lambda}, \mathbf{z})$ is decreasing in z_s, for z_s large enough. Observe that

$$\begin{aligned}\frac{\partial W(\boldsymbol{\lambda}, \mathbf{z})}{\partial z_s} =& \lambda_s (V_s'(z_s) - C_0'(Z)(T_s(z_s)z_s)') \\ \leq& \lambda_s (V_s'(z_s) - C_0'(0)(T_s(z_s)z_s)') \,.\end{aligned}$$

Since V_s is bounded and increasing, then $V_s' \to 0$ as $z_s \to \infty$, while our assumption that $T_s(+\infty) > 0$ implies that $\liminf_{z_s \to \infty} (T_s(z_s)z_s)' > 0$. Therefore, there exists some $\tilde{z}_s$ so that $\frac{\partial W}{\partial z_s} < 0$ for $z_s > \tilde{z}_s$, independently of other variables. This immediately implies that the supremum of $W(\boldsymbol{\lambda}, \mathbf{z})$ is attained for $\{z_s \leq \tilde{z}_s\}$. But this defines a compact region and $W(\boldsymbol{\lambda}, \mathbf{z})$ is continuous, so that the maximum is attained there. □

This leads us to the main result of this section.

THEOREM 10. *There exists a unique solution* $\{\lambda_s^*, z_s^*\}$ *to the social optimization problem* (23). *This optimal solution is internal (*$0 < \lambda_s^* < \lambda_s^{\max}$, $z_s^* > 0$*) and obeys the first order conditions* (25)-(24).

PROOF. By the last lemma, the maximum is attained at a finite point. But Lemmas 6 and 8 that there exists at most one local maximum, which is therefore the global maximum. The second part follows from Lemma 6. □

5. SOCIALLY-OPTIMAL PRICING

Having identified the socially optimal solution, we are faced with the task of implementing this solution. Ideally, such an implementation should not allow the central controller access to private information of the users, which in particular includes their service utility and preferences. In this section we show that the simple per-unit pricing mechanism, with the same price to all, suffices to induce the social optimum.

Let $\{\lambda_s^*, z_s^*\}$ be the unique socially-optimal solution (23). We set the per-unit price to be

$$P = C_0' \left(\sum_s \lambda_s^* z_s^* T_s(z_s^*) \right) \triangleq P^* \,. \tag{27}$$

Recall that each user maximizes his individual utility given this price, as described in Section 3. The main result of this paper is the following one.

THEOREM 11. *Let the per-unit price be* P^**, as defined in* (27). *Then individual optimality leads to the socially optimal solution* $\{\lambda_s^*, z_s^*\}$.

PROOF. We will show that the individual optimality conditions coincide with the conditions for social optimum. Let $\{\lambda_s, z_s\}$ denote the arrival rate and resource allocations that are obtained through individual optimality with price P^*. Observe that z_s is uniquely determined by equation (18), whereas λ_s is given by (19). Comparing equations (24)–(26) with equations (27), (18) and (19), it may be seen that both $\{\lambda_s^*, z_s^*\}$ and $\{\lambda_s, z_s\}$ satisfy equations (25)–(26), with $\eta = P^*$. But by Lemma 7 the solution to these equations is unique, so that $\{\lambda_s, z_s\} = \{\lambda_s^*, z_s^*\}$. □

We next consider the social welfare as a function of the price, and establish its unimodality. Besides its own interest, this property will also be useful below.

PROPOSITION 12. *Let* $W(P)$ *denote the social welfare* $W(\boldsymbol{\lambda}, \mathbf{z})$ *obtained under price* P. *Then* $W(P)$ *is strictly increasing in* P *for* $P < P^*$, *and strictly decreasing for* $P > P^*$.

PROOF. Differentiating $W(P)$ from (22), we obtain

$$\begin{aligned}\frac{dW(P)}{dP} =& \sum_s \frac{\partial W(\boldsymbol{\lambda}, \mathbf{z})}{\partial \lambda_s} \frac{d\lambda_s}{dP} + \frac{\partial W(\boldsymbol{\lambda}, \mathbf{z})}{\partial z_s} \frac{dz_s}{dP} \\ =& \sum_s \bar{V}_s'(\lambda_s) + V_s(z_s) - T_s(z_s) z_s C_0'(Z) \ \frac{d\lambda_s}{dP} \\ &+ \sum_s \lambda_s V_s'(z_s) - \lambda_s (z_s T_s(z_s))' C_0'(Z) \ \frac{dz_s}{dP} \,.\end{aligned}$$

Observing (18) and (19), this gives after some calculation (which we omit here)

$$\frac{dW(P)}{dP} = (P - C_0'(Z)) \frac{dZ}{dP} \,.$$

Now, $Z = \sum_s \lambda_s z_s T_s(z_s)$ is decreasing in P by Lemma 4. Thus, $C_0'(Z)$ is decreasing in P, while the equality $P = C_0'(Z)$ holds at P^*. Therefore $P - C_0'(Z) < 0$ for $P < P^*$ and $P - C_0'(Z) > 0$ for $P > P^*$. This induces opposite signs for $\frac{dW}{dP}$. □

6. ECONOMIC CONTEXT

The cloud computing environment examined in this paper is that of a dynamic service system with sequential arrivals of users, and variable service time that depends on the user choices. The main issues examined, evolving around the notion of social welfare and its maximization, are fundamental ones in microeconomic theory. It will thus be useful to elaborate further on the economic context, and compare the standard models with the ones considered here.

The economic setup of this paper is basically that of a monopoly, namely a single firm that can set market prices. The textbook version of this problem [18], restricted to a single continuous product, considers a finite set I of consumers, with $v_i(x_i)$ denoting the value of consumer i for consuming quantity $x_i \geq 0$ of the product, and $C(x)$ being the cost of production of quantity x. The social welfare is therefore $W(\mathbf{x}) = \sum_i v(x_i) - C(\sum_i x_i)$. With linear pricing, each user is maximizing $V_i(x_i) - Px_i$, and (under standard convexity assumptions) the social optimum is defined by marginal cost pricing, so that $P = C'(\sum_i x_i)$ holds at the optimal point.

Comparing with (3) and (21), it may be seen that the the role of the quantity x_i is taken up in our model by the quantity-time multiple $z_i T_i$. That is, the product being offered here is not measured in terms of the resource quantity z_i itself, but rather in terms of quantity multiplied by usage time. And indeed, the proposed pricing scheme (and the socially optimal one in particular) are linear in the latter measure. This is of course quite reasonable; however, it is important to realize that this structure is not assumed a-priori, but rather arises out of our model once we determine that that operation cost C_0 is a function of the average load, and employ Little's law to describe the effect of demand on the system load. We will briefly comment on other possibilities below.

Let us consider further the use of $x = zT(z)$ (with the type index removed for convenience) as the basic decision variable in place of z. Assume for simplicity that $x(z) \stackrel{\triangle}{=} zT(z)$ is strictly increasing in z, so that the inverse $z(x)$ is well defined (this is indeed the case for $T(z) = a + \frac{D}{z}$, as in (5)). The individual utility (3) can now be expressed as a function of x as

$$\tilde{U}(x) \stackrel{\triangle}{=} U(z(x)) = V(z(x)) - Px\,.$$

It is now easy to verify that condition (ii) of Assumption 3 is equivalent to strict convexity of $V(z(x))$ in x. Therefore, under our assumptions, the individual utility function U is convex when considered as a function of x. In the present paper we have chosen to work with z throughout, motivated by cloud applications where the user actually chooses the resources z explicitly, with the computing time $T(z)$ being determined as a result. We note that working with the resource-time multiple x directly, rather than z, may be natural is other applications when resources are automatically adjusted by the manager according to the application needs. We leave further elaboration of this approach for future study.

An important point to make is the relation between the structure of the operating cost term C_0 and the form of the price tariff. In this paper we have assumed that $C_0 = C_0(Z)$ is a function of the average load $Z = \sum_s \lambda_s z_s T_s(z_s)$, which indeed we believe to be the dominant term. Consider, however, the addition of a cost term C_1 which depends only on the average *number of users in the system (rather the resources utilized*, namely $C_1 = C_1(N)$, with $N = \sum_s \lambda_s T_s(z_s)$. This can represent, for example, the accounting overhead associated with each user. Then, using similar reasoning as before, we are led to consider a two-part tariff of the form $PzT + QT$. We conjecture that a proper choice of the price coefficients P, Q will lead the system to social optimality; again, this is left for further study.

Finally, we comment on our assumption of a continuum of user types. As noted, the textbook model described above considers a finite population I of users. A variant of the model due to Aumann [2] (and see [6]) considers a continuum of infinitesimal users, so that the social welfare, for example, takes the form $W(\mathbf{x}) = \int_i v(x_i)m(di) - C(\int_i x_i m(di))$. This is indeed mathematically akin to (21). However, we note that this similarity is only mathematical. In our model, the users are of finite size, and their number is countable; what is assumed continuous is the pool of possible user types, from which the type of each user is drawn. Therefore, what makes the effect of each user negligible is not this miniature size, but rather the consideration of the average system utility over a long (infinite) time horizon. This is again a distinguishing aspect of the dynamic model considered here, as compared with the standard economic setup.

7. CONCLUSION

This paper considered the resource allocation problem in a cloud computing facility, where the underlying objective is to maximize the social utility through a simple pricing scheme. We showed that the socially optimal operating point is unique, and can be sustained by a linear, usage-based tariff, which charges a fixed price per unit resource and unit time.

Besides the analytical results, a major contribution of the paper is in the modeling aspect. The proposed model, which is well suited for economic analysis, incorporates several novel features that pertain to the cloud computing environment, including:

– Incorporating temporal aspects into the model.

–Flexible dependence of the computation time on the applied resources, which can be used to take account of setup and parallelization overheads.

– User heterogeneity, in terms of both utility and job processing requirements.

– A flexible finite-class, continuous-type model that allows smooth demand functions along with finite-dimensional problem formulation.

– Variable arrival rates, which is shaped by user balking, in addition to their choice of resources.

The essential model developed here may provide a basis for additional work on economic aspects of cloud computing, considering further aspects of revenue, profit and competition among clouds. Interesting extensions to the model include the allocation of multiple resource types, or resource bundles, rather than the single resource type considered here, as well as the consideration of discrete resources, and more a detailed analysis of resources allocations that are time-varying according to the application's needs. Finally, it should be of interest to study possible effects of congestion, which were assumed here to be negligible due to proper management. We hope that the model presented in this paper will provide a convenient starting point to study these important problems.

8. REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

[2] R. J. Aumann. Markets with a continuum of traders. *Econometrica*, 32:39–50, 1964.

[3] V. S. Borkar. *Stochastic Approximation: A Dynamic Systems Viewpoint*. Hindustan Book Agency, New Delhi, 2008.

[4] R. Buyya, D. Abramson, and S.Venugopal. The grid economy. *Proceedings of the IEEE*, 93(3):698–714, 2005.

[5] C. Courcoubetis and R. Weber. *Pricing Communication Networks: Economics, Technology and Modelling*. Wiley, Chichester, England, 2003.

[6] B. Ellickson. *Competitive Equilibrium: Theory and Applications*. Cambridge University Press, Cambridge, UK, 1993.

[7] I. Foster, C. Kesselman, and S. Tuecker. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of High Performance Computing Applications*, 15(3):200–222, 2001.

[8] R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Springer, 2003.

[9] M. D. Hill and M. R. Marty. Amdahl's law in the multicore era. *IEEE COMPUTER*, 41:33–38, 2008.

[10] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, second edition, 2003.

[11] A. Mas-Colell, M. Whinston, and J. Green. *Microeconomic theory*. Oxford University Press New York, 1995.

[12] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Oper. Res.*, 38(5):870–883, 1990.

[13] B. M. Mitchell and I. Vogelsang. *Telecommunications Pricing: Theory and Practice*. Cambridge University Press, Cambridge, England, 1991.

[14] National Institute of Standards and Technlogy. NIST Definition of Cloud Computing v15. http://csrc.nist.gov/groups/SNS/cloud-computing, July 2009.

[15] K. W. Ross. *Multiservice Loss Models for Broadbank Telecimmunication Networks*. Springer, London, 1995.

[16] S. Shakkottai and R. Srikant. Network optimization and control. In *Foundations and Trends in Networking*, volume 2, pages 271–379. Now Publishers Inc, 2008.

[17] S. Stidham. A last word on $l = \lambda w$. *Operations Research*, 22(2):417–421, 1974.

[18] H. R. Varian. *Microeconomic Analysis*. Norton and Company, 1992.

[19] C. S. Yeo and B. Rajkumar. A taxonomy of market-based resource management systems for utility-driven cluster computing. *Software Practice and Experience*, 36(13):1381–1419, 2006.

[20] Y. Yi and M. Chiang. Stochastic network utility maximization – a tribute to Kelly's paper published in this journal a decade ago. *European Transactions on Telecommunications*, 19(4):421–442, 2008.

[21] B. Yolken and N. Bambos. Game based capacity allocation for utility computing environments. In *ValueTools '08: Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools*, pages 1–8, 2008. To appear in *Telecommunication Systems*.