# Pricing of Service in Clouds: Optimal Response and Strategic Interactions

*(Extended Abstract)*

Bhuvan Urgaonkar, George Kesidis, Uday V. Shanbhag, *and* Cheng Wang
The Pennsylvania State University

## 1. INTRODUCTION

Cloud providers and consumers encounter a variety of pricing models devised to disincentivize the occurrence of high or unpredictable demands (especially simultaneously from multiple consumers) and incentivize certain kinds of behavior (e.g., predictable resource procurement, adhering to requested resource demands, etc.). Examples of such arrangements may have a cloud in the role of (i) a consumer (e.g., a data center procuring power from an electric utility based on tiered [10] or peak-based [5] pricing (we elaborate on both of these momentarily), or network bandwidth from an Internet Service Provider (ISP) based on a high percentile of bandwidth usage [1]), (ii) a provider (e.g., IT customers procuring virtual machines (VMs) or storage from a cloud provider such as Amazon EC2 [3, 2]), or (iii) both (e.g., a Content Distribution Network (CDN) provider such as Akamai renting rack space from another data center and paying the data center rather than an electric utility for provisioned power [9]). Table 1 presents some such situations.

| Utility provider | Consumer | Resource(s) | Pricing scheme |
|---|---|---|---|
| Electric,<br>ISP | Cloud<br>data center | Power,<br>Network b/w | Tiered, peak+avg.,<br>Peak (high %-ile) |
| Cloud<br>data center | CDN | Space+servers,<br>power | Capacity-based,<br>Peak-based |
| IaaS Cloud<br>data center | IT customers | VMs/storage | On-demand<br>(fixed or spot),<br>Reservation-based |

**Table 1:** Examples of cloud data centers or their customers negotiating complex tariffs such as those involving tiered or peak-based pricing.

Each of these situations presents two complementary control and optimization problems. First, once a pricing scheme has been negotiated, *how should the consumer modulate its demand to optimize its profits?* Problems of this kind have received a lot of attention in recent literature: examples include minimizing the electricity bill for time-varying prices [11, 12], minimizing the peak power draw using batteries [4], cost-optimal procurement of VMs from clouds that offer both on-demand (more expensive) or reservation-based "bulk" (cheaper) VMs [13], and optimizing the electricity bill when the pricing scheme charges higher during "coincident" peaks [7], among others. The second problem, which we study in this paper, asks: *how should the provider and consumer negotiate the specific pricing structure they will employ?*

We explore this question by focusing on two popular pricing mechanisms:

- *Tiered*, wherein some demand/usage thresholds define differential per unit prices for the resource(s); a special case has only one threshold, upon exceeding which a high rate (or additionally a "penalty") is imposed (e.g., electric utilities often employ tiered pricing [10]).
- *Demand "tail"-based*, wherein (potentially in addition to average resource usage) charges are applied based on some property of the tail of observed or requested demand distribution (e.g., 95th percentile [1] or peak [5]).

There are other pricing schemes such as "bulk" purchase which are equally important; we do not discuss these in this paper [2, 13] but we do consider them insteresting directions for future work. We sketch initial ideas for decision-making regarding the cost-optimal pricing option at a single consumer in Section 2 followed by strategic interactions between providers and consumers in Section 3. We conclude with directions for future research in Section 4.

## 2. SINGLE CONSUMER

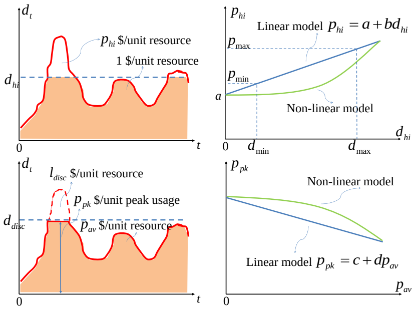

**Figure 1:** Pricing options presented by the provider for our tiered and peak+avg.-based models.

Consider a single consumer $\mathcal{C}$ that procures a resource from a single provider $\mathcal{P}$ according to a pricing scheme they agree upon periodically (e.g., once every few months) and which is then *fixed* for the remainder of this period. Our interest is in the decision-making at $\mathcal{C}$ when choosing a pricing option out of a set of options that $\mathcal{P}$ presents it with. One intuitively appealing goal for $\mathcal{C}$ would be to pick a pricing option that maximizes its *expected profit* (revenue minus costs) over the billing cycle (given knowledge of its demand and control options). We consider a time-slotted model where a billing cycle consists of $T$ control windows. Suppose the consumer's resource demand in window $t$ $(1 \leq t \leq T)$ is denoted as $d_t$. Suppose that $\mathcal{C}$ earns revenue $\rho$ for every unit of resource it procures from $\mathcal{P}$. We explore this decision-making via two case studies with simplifying assumptions about $\mathcal{C}$'s demand, and then discuss ideas for generalizing it in Section 2.3.

## 2.1 Inelastic Demand, Tiered Pricing

For our first case study, we assume a tiered pricing model which has two parameters $d_{hi}$ and $p_{hi}$ with the following meaning: if $d_t \leq d_{hi}$, the per unit resource price is \$1, while if $d_t > d_{hi}$ the per unit resource price is $p_{hi} >$ \$1. We assume that $\mathcal{C}$ poses *inelastic* demand, i.e., it can not modulate its demand in any way and must procure all of $d_t$ from $\mathcal{P}$ during $t$, $\forall t$. Suppose that $\mathcal{P}$ presents pricing options as in Figure 1(a), e.g., by expressing allowed pairs of $(p_{hi}, d_{hi})$ in the form of a function of $p_{hi} = f_{tier}(d_{hi})$. Suppose that $d_t$ is an i.i.d. random variable following the Pareto distribution:

$$f_{d_t}(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & \text{for } x \geq x_m, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $x_m > 0$. $\mathcal{C}$ would pick $(p_{hi}, d_{hi})$ to maximize its expected profit per control window, resulting in the following optimization problem:

$$\max_{p_{hi}, d_{hi}} \rho \int_{x_m}^{\infty} x \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx - \Big\{ \int_{x_m}^{d_{hi}} x \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx + \int_{d_{hi}}^{\infty} \Big( d_{hi} + p_{hi}(x - d_{hi}) \Big) \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx \Big\}.$$

Subject to: $d_{min} \leq d_{hi} \leq d_{max}, 1 < p_{min} \leq p_{hi} \leq p_{max}$.

Since we assume that $\mathcal{C}$ does not modulate its demand (unlike in our subsequent models), this is equivalent to:

$$\min_{p_{hi}, d_{hi}} \int_{x_m}^{d_{hi}} x \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx \; + \int_{d_{hi}}^{\infty} \Big( d_{hi} \; + p_{hi}(x - d_{hi}) \Big) \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx,$$

with the same constraints as before. The objective above simplifies to: $\min_{p_{hi}, d_{hi}} \frac{1}{\alpha - 1} \Big( \alpha x_m + (p-1) x_m^\alpha d_{hi}^{1-\alpha} \Big)$. We can now employ the known function $f_{tier}(.)$ to express the above solely in terms of $d_{hi}$ and derive $d_{hi}^*$ that minimizes the resulting expression. As a specific example, if $p_{hi} = a + b d_{hi}$ ("linear" in Figure 1(a)), we find:

$$d_{hi}^* = \frac{(1-a)(\alpha - 1)}{b(\alpha - 2)}.$$

In Table 2 (row labeled "Inelastic") we present some simple results based on our analysis. As expected intuitively, with higher demand variance, we prefer higher $d_{hi}$.

## 2.2 Elastic Demand, Peak-Based Pricing

Now suppose that $\mathcal{P}$ employs an avg+peak-based pricing scheme with parameters $p_{av}$ and $p_{pk}$ denoting the per unit resource price and the peak usage price per billing cycle, respectively. We continue to assume that $d_t$ follows the Pareto distribution above, which implies that the consumer must cap its demand to a finite value. We assume that it employs a "demand discarding" knob with associated loss of revenue ($l_{disc}$) for every unit of resource demand discarded. $\mathcal{P}$ presents pricing options as in Figure 1(b), e.g., by expressing allowed pairs of $(p_{av}, p_{pk})$ in the form of a function of $p_{pk} = f_{pk}(p_{av})$.

| Demand | Params. | Low var.<br>($\alpha = 100$) | Med. var.<br>($\alpha = 50$) | High var.<br>($\alpha = 2.5$) |
|---|---|---|---|---|
| Inelastic | $(p_{hi}, d_{hi})$ | (1.0092, 181.84) | (1.0188, 183.75) | (2.8, 540) |
| Elastic<br>(discarding) | $(p_{av}, p_{pk})$<br>$d_{disc}$ | (0.05, 10)<br>107.54 | (0.05, 10)<br>115.65 | (1, 105)<br>701.39 |

**Table 2:** **Some results from our case-studies. We choose demand distributions with the same average but different variances for which different pricing options are chosen by our analysis. $x_m$=100, $T$=720, $\rho$=10; options for tiered pricing: $p_{hi} = 0.1 + 0.005 d_{hi}$; options for peak+avg.-based pricing: $p_{pk} = 5 + 100 p_{av}$, $l_{disc}$=10.**

$\mathcal{C}$ would want to choose $p_{av}$, $p_{pk}$, and a demand threshold $d_{disc}$ (any demand in excess of $d_{disc}$ will be discarded) to maximize its expected profit:

$$\max_{p_{av}, p_{pk}, d_{disc}} - p_{pk} d_{disc} + T \Big\{ \int_{x_m}^{d_{disc}} (\rho - p_{av}) x \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx + \int_{d_{disc}}^{\infty} \Big( (\rho - p_{av}) d_{disc} - l_{disc}(x - d_{disc}) \Big) \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx \Big\},$$

which corresponds to:

$$d_{disc}^* = x_m \Big( \frac{T(\rho + l_{disc} - p_{av})}{p_{pk}} \Big)^{1/\alpha}.$$

In Table 2 (row labeled "Elastic") we present some simple results based on the above analysis. As expected intuitively, $d_{disc}$ increases with the variance in demand.

## 2.3 General Elastic Demand

As an initial attempt at dealing with general demands and control options, we present a simple model for optimization based on chance (in-probability) constraints, and leave for future work more complex models leveraging the large literature on queues controlled by Markovian decisions. Let $\mu$ be the mean of the consumer's "raw" resource demand $\Delta$. We consider more general demand modulation (based on delaying part of the demand) wherein units of demand will drop out and not return if they are deferred (initially delayed) by more than $\tau$ seconds. In addition to $\mu, \tau$, suppose $\Delta$ is also modeled by a "generalized stochastically bounded burstiness" curve [15] $\{(r, \phi(r\tau)) \mid r > \mu\}$; i.e., a queue whose arrivals are $\Delta$ and is served at rate $r$ will have backlog $Q_r$ such that

$$\mathsf{P}(Q_r \geq r\tau) \leq \phi(r\tau).$$

We would explore ways of learning the demand *model* $(\mu, \tau, \phi(\cdot))$; significant literature exists on this. For example, this could be based on analysis of day-ahead prices of power or more extensive historical demand data recorded in the same context as the present demand, $\Delta$, under consideration). As another example, jobs/requests arriving at a data center could be classified so that their expected virtual machine needs or energy requirements, and $Q_r$ in turn, could be estimated.

Note that the *peak rate* of the queue-output is limited to $r$. Assume that the load shed by this regulating queue is $\mu\phi(r\tau)$; this is an approximation in particular because the arrivals are not Poisson so that PASTA [14] does not apply. Our *offline* objective is to select $r$ to maximize the net revenue of the data center modeled as $(\rho - c) T \mu (1 - \phi(r\tau)) - f(r)$, where $\rho > c$ respectively are the revenue and cost.The resulting optimization problem is given by

$$\max_{r \in \mathcal{R}} h(r) \triangleq (\rho - c) T \mu (1 - \phi(r\tau)) - f(r),$$

where $\mathcal{R} \triangleq [0, \bar{r}]$ and $\bar{r}$ denotes the upper bound on the rate. Under the assumption that $f$ and $\phi$ are continuously differentiable and convex functions of $r$, the optimal rate $r^*$ is given by a solution to a variational inequality problem $\mathrm{VI}(\mathcal{R}, -\nabla_r h)$ [6]. It may be recalled that $r^*$ is a solution of $\mathrm{VI}(\mathcal{R}, -\nabla_r h)$ if

$$(r - r^*)^T \nabla_r h(r) \leq 0, \quad \forall r \in \mathcal{R}.$$

An interesting direction for future work would explore if a formulation such as above is general enough to allow the ones in Sections 2.1 and 2.2 and others (or appropriate approximations of these) as its special cases.

# 3. STRATEGIC INTERACTIONS

A natural concern in a multi-agent setting lies in the impact of competitive interactions. In this section, we consider several variants of such interactions. We begin by examining a noncooperative Nash game played amongst a collection of consumers faced by a single non-strategic provider. Subsequently, we allow for the provider to assume a leadership role in designing pricing structures while being cognizant of the Nash equilibrium that ensues between the consumers.

## 3.1 Strategic Consumers

Consider a noncooperative setting comprising of $N$ consumers facing a fixed pricing structure. Suppose consumer $i$'s profit function is denote by $h_i(r_i; r_{-i})$ where

$$h_i(r_i; r_{-i}) \triangleq (\rho_i - c_i)T\mu(1 - \phi(\tau r_i)) - f\left(\sum_{i=1}^{N} r_i\right),$$

$\rho_i$ denotes consumer $i$'s revenue, $c_i$ denotes her cost, and $r_{-i} = (r_j)_{j \neq i}$. The function $f$ denotes a continuously differentiable increasing convex cost function associated with the coincident peak, given $\sum_{i=1}^{N} r_i$. The resulting Nash equilibrium problem is given by a tuple $\{r_i^*\}_{i=1}^{N}$ where for $i = 1, \ldots, N$, $r_i^*$ is a solution of the following parameterized problem:

$$r_i^* \in \operatorname{argmax}_{r_i \in \mathcal{R}_i} \; h_i(r_i; r_{-i}), \qquad (\mathrm{Prob}(r_{-i}))$$

Under suitable convexity assumptions, $r^*$ is a Nash equilibrium if and only if it is a solution to $\mathrm{VI}(\mathcal{R}, F(r))$ where

$$\mathcal{R} \triangleq \prod_{i=1}^{N} \mathcal{R}_i \text{ and } F(r) \triangleq \begin{pmatrix} -\nabla_{r_1} h_1 \\ \vdots \\ -\nabla_{r_N} h_N \end{pmatrix}.$$

Proposition 1. *Consider a Nash game in which the ith consumer solves* $(Prob(r_{-i}))$ *for* $i = 1, \ldots, N$. *Then a unique Nash equilibrium exists.*

Note that existence follows from the nonemptiness and compactness of $\mathcal{R}$ and the continuity of $F(r)$. Furthermore, by observing that the map $F(r)$ is strictly monotone over $\mathcal{R}$ and noting that a solution exists, it follows that $\mathrm{VI}(\mathcal{R}, F(r))$ admits a unique solution. A related question is whether this equilibrium is efficient with respect to a problem where all users are centrally controlled. If not, are there choices of price designs that lead to minimal efficiency loss from the standpoint of system welfare.

## 3.2 Strategic Providers

One avenue for designing pricing structures is to consider a single-leader multi-follower problem in which the service provider takes on the garb of a leader while the followers are the consumers. Through such a model, the service provider optimizes the choice of pricing parameters over the set of equilibria arising from the Nash interactions between the followers. For instance, consider a problem where the provider (leader) chooses $\mu$ and $\tau$ and the consumers (followers) reach an equilibrium by responding to this choice:

$$\max_{r \in \mathcal{R}, \tau, \mu} \; w(\mu, \tau, r)$$
$$\text{Subject to } \; r \text{ solves } \mathrm{VI}(\mathcal{R}, F(r, \theta)),$$

where $w(\mu, \tau, r)$ denotes the welfare function of the service provider. This problem is known as a mathematical program with equilibrium constraints (MPEC) [8], a class of nonconvex programs, but given that the Nash equilibrium problem is defined via a strictly monotone map, we believe that an *implicit* form may be constructed.

$$\max_{\tau, \mu} w(\mu, \tau, r(\mu, \tau)),$$

where $r(\mu, \tau)$ denotes a single-valued map that defines the Nash equilibrium, given $\tau$ and $\mu$. Note that the claim of single-valuedness is by no means easy to make but rests on the structural properties of the map.

A final question is whether one may consider a problem where a collection of providers compete for business across a set of consumers. In such an instance, the resulting problem reduces to a *multi-leader multi-follower game* in which the providers compete in a Nash game in which each provider solves an MPEC.

# 4. CONCLUSIONS

Whereas price negotiation is a very general concern, the unique and idiosyncratic properties of cloud workloads, their performance needs, and the impact of workload modulation on their resource consumption and revenue imply that this area offers rich and novel modeling/optimization problems. In particular, cloud environments appear to present us with collections of different/hybrid pricing schemes among the same set of providers and consumers (e.g., tiered pricing for power and tail-based pricing for network bandwidth faced by CDNs [9]). We plan to enhance our modeling to capture these scenarios as part of our future work.

# 5. REFERENCES

[1] M. Adler, R. K. Sitaraman, and H. Venkataramani. Algorithms for optimizing the bandwidth cost of content delivery. *Computer Networks*, 55(18):4007–4020, 2011.

[2] Amazon EC2 Reserved Instances. http://aws.amazon.com/ec2/reserved-instances/.

[3] Amazon EC2 Spot Instances. http://aws.amazon.com/ec2/spot-instances/.

[4] A. Bar-Noy, M. Johnson, and O. Liu. Peak shaving through resource buffering. In *WAOA*, 2008.

[5] Duke Energy Carolinas, LLC: SCHEDULE LGS (NC) LARGE GENERAL SERVICE, date effective: 02/01/2013, 2013. http://www.duke-energy.com/pdfs/NCScheduleLGS.pdf.

[6] F. Facchinei and J.-S. Pang. *Finite Dimensional Variational Inequalities and Complementarity Problems: Vols I and II*. Springer-Verlag, NY, Inc., 2003.

[7] Z. Liu, A. Wierman, Y. Chen, and B. Razon. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. In *ACM SIGMETRICS*, Jun 2013.

[8] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, 1996.

[9] E. Nygren, R. Sitaraman, and J. Sun. The akamai network: A platform for high-performance internet applications. *ACM SIGOPS Operating Systems Review*, 44(3):2–19, 2010.

[10] F. Taylor-Hochberg. Pricing of electricity could use a jolt, Op-Ed, LA Times, Nov. 13, 2011. http://articles.latimes.com/2011/nov/13/opinion/la-oe-taylor-hochberg-electricity-rates-20111113.

[11] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam. Optimal power cost management using stored energy in data centers. In *ACM SIGMETRICS*, pages 221–232, Jun 2011.

[12] P. M. van de Ven, N. Hegde, L. Massoulié, and T. Salonidis. Optimal control of end-user energy storage. *CoRR*, abs/1203.1891, 2012.

[13] W. Wang, B. Li, and B. Liang. To reserve or not to reserve: Optimal cost management for iaas clouds. In *USENIX ICAC*, 2013.

[14] R. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.

[15] Q. Yin, Y. Jiang, S. Jiang, and P. Kong. Analysis of generalized stochastically bounded bursty traffic for communication networks. In *Proc. IEEE LCN*, 2002.