Competition on Price and Quality in Cloud Computing

Cinar Kilcioglu
Columbia University
Graduate School of Business
ckilcioglu16@gsb.columbia.edu

Justin M. Rao
Microsoft Research
Redmond, WA
justin.rao@microsoft.com

ABSTRACT

The public cloud "infrastructure as a service" market possesses unique features that make it difficult to predict long-run economic behavior. On the one hand, major providers buy their hardware from the same manufacturers, operate in similar locations and offer a similar menu of products. On the other hand, the competitors use different proprietary "fabric" to manage virtualization, resource allocation and data transfer. The menus offered by each provider involve a discrete number of choices (virtual machine sizes) and allow providers to locate in different parts of the price-quality space. We document this differentiation empirically by running benchmarking tests. This allows us to calibrate a model of firm technology. Firm technology is an input into our theoretical model of price-quality competition. The monopoly case highlights the importance of competition in blocking "bad equilibrium" where performance is intentionally slowed down or options are unduly limited. In duopoly, price competition is fierce, but prices do not converge to the same level because of price-quality differentiation. The model helps explain market trends, such the healthy operating profit margin recently reported by Amazon Web Services. Our empirically calibrated model helps not only explain price cutting behavior but also how providers can manage a profit despite predictions that the market "should be" totally commoditized.

Keywords

Game theory; cloud computing; pricing

1. INTRODUCTION

Cloud computing uses two key pieces of technology. The first is virtualization, the ability to create a simulated environment that can run software just like a physical computer. Virtualization is governed by the "cloud fabric," which functions as the hypervisor, scheduler and manages fault tolerance. The second piece is network communication protocol, both within the datacenter and between different datacenters. While both technologies have been around for decades, there have been many proprietary advances and thus the quality of the service offered can vary, even when datacenters use the same underlying physical hardware. An analogy is the operating system of a single computer—firms invest in operating system technology to improve performance given the expected capabilities of the underlying hardware.

While the importance of these technologies is widely researched in the systems community (see *e.g.*, [21, 23]), the "infrastructure as a service" public cloud marketplace is often described as "commoditized" from an economic competition perspective [17]. The reasoning is putatively straightforward. Since cloud providers use similar, if not identical, physical hardware they cannot meaningfully differentiate their products and thus profit margins should converge to zero. We begin our analysis by empirically assessing this claim by running a series of benchmarking workloads across two major provider's various service levels ("virtual machine" (VM) size), similar to the approach used in [16]. We find different run-times for similarly described offerings, such as "2 virtual cores, 4GB memory." While runtime decreases for both providers as one moves to larger VMs, the price-performance trade-offs are different, which means there are different feasible price-quality combinations. We formalize this insight with a two-parameter model of the firm's production technology and the calibrated model achieves good fit to our data.

The fitted parameters are used in our theoretical model as one source of differentiation across firms. We view these technologies as fixed for our analysis, imagining they are the result of countless engineering decisions made over the years. Endowed with a technology, firms then choose *performance menus*, which provides a second source of potential differentiation. A performance menu is a set of VMs with different CPU, memory and disk configurations. For example, Amazon Web Services (AWS) offers about 20 different VM configurations, ranging from low performance "micro" to high performance "extra large." We model customers as having heterogeneous types with varying sensitivity to job completion time, but with a common job completion valuation and workload requirement. Customers choose optimally from the price-quality menus provided by firms.

We start with the monopoly case. There are a number of reasons this is a useful starting point even though most large regional markets are not currently characterized by monopoly. First, SEC Filings reveal that AWS is currently many times larger than the next closest competitor, indi-

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883043.

cating that one provider "pulling away" from the competition is certainly not implausible. Second, smaller countries often have only a single major provider with a datacenter within national boundaries. Finally, a customer that has used a given provider for some time could face large switching costs, leading to potential monopolistic dynamics targeted at "locked in" customers.

For the monopoly case, we characterize the optimal base price, quality level and associated customer demand functions. Interestingly, under some conditions, offering an additional quality level does not generate more revenue. We provide sufficient conditions for when a firm should offer multiple quality levels. The conditions show that when the quality level is increasing almost linearly in price and there are some customer types in the system that are highly sensitive to delay, offering an additional higher quality products, up to a point, generates more revenue.

The results also reveal an interesting dynamic with respect to customer valuations and quality. When valuations increase, the optimal strategy for the service provider is to intentionally degrade the quality level of lower tier offerings as opposed to increasing the unit price. While this might sound counter-intuitive at first, it is readily understood by recognizing that customers are paying per time-unit. A higher quality product is not only more expensive, but offers faster runtime—the faster runtime reduces the net payment on the margin. As valuations increase, there is an incentive to make the low quality options less attractive to "high types." By damaging the product, it is effectively more expensive *and* less attractive due to increased delay. This "double dividend" for damaging the good has previously been observed in the computing hardware and shipping/transport industries [13]. Overall, the results for the monopoly case highlight the nuanced role of competition in this marketplace.

We next move on to the duopoly case. We start by characterizing the Nash equilibrium when each service provider is restricted to offer only one quality level. In this case the higher quality provider attracts high-type customers (the ones that are more sensitive runtime delays) at a higher price. In other words, there is stable differentiation on the quality dimension. When providers are allowed to offer multiple quality levels, we no longer have a closed form solution. We thus simulate the game under different market settings where providers compete in base price level. Interestingly, prices do not converge and instead display Edgeworth cycles (as in [18]). The intuition for these cycles is the standard one, with a a bit of tweak. Despite the quality differentiation, the goods are relatively good substitutes for each other and thus Bertrand-like price competition leads to successive undercutting of price, albeit at different price levels (the tweak). That is, prices move in parallel down to a point of very low returns for the firm. At this point, a war of attrition ensues and one firm "leads" the pair back up to a higher price point and the cycle repeats.

Past research has shown that these types of cycles, though commonly predicted, are empirically quite rare. Exceptions occur in markets where prices change flexibly and there are other sources of price volatility (e.g. due to cost shifters such as oil prices in the retail gasoline market [20]). Perhaps unsurprisingly, then, we do not observe classical Edgeworth cycles in cloud computing. It turns out, however, that once we consider important market features the observed price patterns share qualitatively similar features with classic cycles.

The most important dynamic is the relatively rapid reductions in the cost per compute cycle due to technological advances, which are commonly attributed to Moore's law. In reality the situation is more complex, with Moore's law slowly giving way and other advances breaking through [12]. Nonetheless, these advances provide both a real decline in costs for the provider and a strong consumer perception that prices should fall, not rise. In practice, cloud providers tend to replace physical hardware approximately every three years. The release of new hardware enables new, superior "generations" of VMs. But the old generations can nonetheless be virtualized on the new hardware, just with less physical resources required than before and thus at a lower cost. This means constant prices for older generations are effective increases relative to costs. We examine historical prices and observe that the largest provider, AWS, tends to offer newer generations at lower prices and keep older generation prices relatively high. Indeed we document that older, inferior generations are often priced *higher* than the comparable VMs in the new generations. So while the model predicts varying intensity of price competition over time, in practice we observe this variance across products by release date. In other words, some "regions" of the product space have vigorous competition—we view this as substantively similar to the cycling prediction.

Further, we highlight that our model predicts that price differences can be maintained in equilibrium and the market will not totally commoditized. Interestingly, in the Summer of 2015 one major provider dropped prices rather substantially and the other two major providers did not follow suit. Our model gives a rigorous explanation as to why.

Related Literature. To the best of our knowledge, it is the first paper that models the cloud computing products from price-quality perspective under competition. The analysis draws on three main streams of literature. The first is from economics and marketing literature on price-quality competition. Most papers here focus on the case when players are symmetric and each player chooses one quality and one price under competition or one player chooses two distinct quality levels under monopoly ([18, 19, 25]). Here we have two asymmetric players each choosing multiple quality levels and prices, and both quality levels and prices are interdependent, which is why we have to rely on simulations at times.

The second stream of literature is on cloud pricing. [8, 27] look at the problem from a higher level and try to find the best pricing strategy by offering the same product in different pricing mechanisms. In this work, we aim to find a revenue maximizing price-quality menu with fixed prices. There are papers on competition in an oligopoly market with multiple providers. [14] studies non-cooperative competition model in a cloud market and computes an equilibrium price. However, each player has single product type in this study. [10] studies the price competition in cloud computing by considering all three layers of cloud. Our focus in this study is only the IaaS market.

The third stream is the analysis reports prepared by private cloud companies ([2, 3, 7]). They investigate the performance of different cloud providers from different angles. Although their methodology contains extensive performance analysis, it does not have a solid economic framework, and

performance values and units prices are not incorporated into the analysis in a transparent manner.

2. MODEL

On the customer side, there are n customer types indexed by i, where customer type i has a valuation (v_i), delay sensitivity (c_i), both per unit time of workload[1] under nominal quality level, and arrival rate (λ_i). We assume there is only one type of workload which can be parallelizable up to a certain extent, and all customer types need to run the same workload. We relax this assumption and discuss the results in §5.

On the provider side, there are m different service providers indexed by j, where service provider j chooses a base quality level q_{j1} ($0 < q_{j1} < \bar{q}_j$), where $\bar{q}_j$ is the maximum base quality level that can be offered, price per unit time for the base quality level p_{j1}, and number of quality levels to offer L_j. Each service provider has an inherent performance scaling factor α_j determined by the structure and technology used (which will later estimate, $0.5 < \alpha_j < 1$), and each offers a price-quality menu (p_{jk}, q_{jk}), where $p_{jk} = 2^{k-1}p_{j1}$, $q_{jk} = 2^{k-1}\alpha_j^{k-1}q_{j1}$ for $k = 1, 2, \ldots, L_j$.

The size of a workload is defined as the time it takes to complete the job using a baseline quality product. We are assuming job completion time function $W(w, q) := \frac{w}{q}$ where w is the completion time of a job under baseline quality and q is the quality level.

The utility of customer type i with workload w, choosing quality level k of service provider j is

$$\begin{aligned} U_{ijk} &= v_i w - c_i W(w, q_{jk}) - p_{jk} W(w, q_{jk}) \\ &= w\left(v_i - \frac{c_i + 2^{k-1}p_{j1}}{2^{k-1}\alpha_j^{k-1}q_{j1}}\right), \end{aligned}$$

with $U_{ij0} = 0$ representing the no-buy option.

Then, customer type i chooses quality level k^* of service provider j^*, where

$$\begin{aligned} j^* &= \underset{j\in\{1,2,\ldots,m\}}{\text{argmax}} \{\max_{k\in\{0,1,2,\ldots,L_j\}} U_{ijk}\} \text{ and} \\ k^* &= \underset{k\in 0,1,2,\ldots,L_{j^*}}{\text{argmax}} U_{ij^*k}. \end{aligned}$$

Service providers are revenue maximizers.[2] Assuming each customer type has workload w, the revenue function for service provider j is

$$\Pi_j(p_{j1}, q_{j1}) = w\left[\sum_{i\in\mathcal{S}_{j1}} \lambda_i \frac{p_{j1}}{q_{j1}} + \sum_{i\in\mathcal{S}_{j2}} \lambda_i \frac{p_{j1}}{\alpha_j q_{j1}} + \ldots + \sum_{i\in\mathcal{S}_{jL_j}} \lambda_i \frac{p_{j1}}{\alpha_j^{L_j-1} q_{j1}}\right],$$

where S_{jk} is the set of customer types that choose quality level k of service provider j ($k = 1, 2, \ldots, L_j$).

Model Validity. All big cloud providers offer different product families to their customers, and each product family is customized for special kind of workloads. Amazon has *t2*, *m4*, *c4*; Google has *standard*, *high-mem*; and Microsoft has *A*, *D*, *G*, to name a few ([1, 5, 6]). In most of these product families companies offer 4 different product sizes with different prices; however, what they actually pick is a base level product configuration and a price for this base level. Once the base level is picked, second product is configured as the double the size of the base product with twice the price, third product is configured as the double of the second product, and finally fourth is configured as the double of the third product. Price - Configuration menu for Microsoft's D product family with Linux Machine for Central US region is given in Table 1 as an example of this structure.

Table 1: Azure Price - Configuration Menu

Product	Cores	Ram	Disk Sizes	Unit Price
D1	1	3.5 GB	50 GB	$0.077/hr
D2	2	7 GB	100 GB	$0.154/hr
D3	4	14 GB	200 GB	$0.308/hr
D4	8	28 GB	400 GB	$0.616/hr

To validate our price-quality model, we have picked two service providers (a and b) with one product family for each. Therefore, we have products a_i and b_i, lower i indicating smaller size product, with unit prices $2^{i-1}0.100$ and $2^{i-1}0.126$ ($i = 1, 2, 3, 4$) for providers a and b, respectively.[3] The workload we have chosen for this experiment is *DaCapo* ([4, 11]). DaCapo is a benchmark suite that runs different Java workloads with non-trivial memory loads. We have run the workload once a day for one week at the same time for both providers with different product sizes in similar regions. Average running times and cost values are summarized in Table 2.[4] Contrary to the previous literature ([22, 24, 26]), our experiment with one type of workload has shown that the job completion time does not vary too much over time for the same product (the average standard deviation in completion time is less than 5% of the mean completion time per product), unless the product is a burstable type product, or has a shared CPU (*t2* product family in AWS, *f1-micro* in Google).

Table 2: Price-Quality Comparison

Product	Unit Price	Avg. Comp. Time	Total Cost
a_1	$0.100/hr	738.14 sec	$0.021
a_2	$0.200/hr	490.47 sec	$0.027
a_3	$0.400/hr	383.90 sec	$0.043
a_4	$0.800/hr	360.57 sec	$0.080
b_1	$0.126/hr	719.71 sec	$0.025
b_2	$0.252/hr	468.00 sec	$0.033
b_3	$0.504/hr	360.71 sec	$0.051
b_4	$1.008/hr	308.71 sec	$0.086

Figure 1 shows how products are located in time/cost space for this specific workload. User utility increases as we move towards the origin, as it signals faster performance and lower cost. Interestingly, all product offerings are Pareto efficient, that is, there is no product that is both cheaper and

[1]We use *workload* and *job* interchangeably throughout the paper.

[2]We later discuss how to incorporate costs in the analysis.

[3]For anonymity, names are filtered and unit prices are transformed.

[4]In total cost calculations, it is assumed that cost is incurred per second basis.

faster than any other products. Therefore, each product can be chosen by a rational customer based on her time/cost trade-off. Since users differ with respect to time sensitivity, they will choose different performance level.

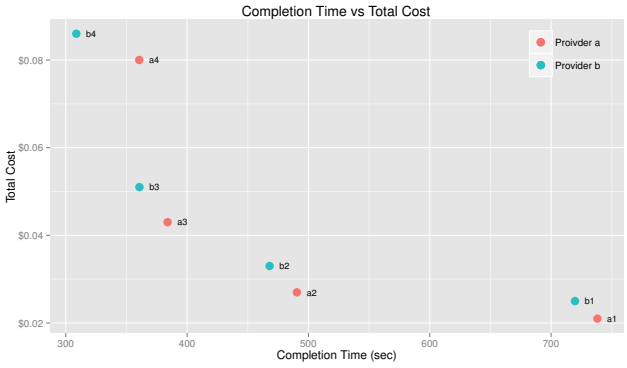


Figure 1: Completion Time vs Total Cost

Assuming $w = 1000$ for the workload we are experimenting with, we try to estimate the scaling factor and base quality level for both products. We find that $(\alpha_1, q_1) = (0.693, 1.355)$, and $(\alpha_2, q_2) = (0.616, 1.733)$ for service providers a and b. respectively. The mean percentage absolute error of our fit is 8% for both providers.[5] Hence, we can conclude that our model with quality level function $2^{k-1}\alpha_j^{k-1}q_{j1}$ is fairly realistic. Note that in reality, α value is not only provider dependent, but also workload dependent. No matter how good infrastructure one provider has, if the workload to be run is not parallelizable, α value would end up being low. We are doing our analysis for a specific type of workload which is fairly parallelizable, and we discuss possible extensions to this in §5. Reader may refer to [9] and [15] for detailed analysis on maximum achievable performance gain with parallelization formulations based on workload type.

3. REVENUE MAXIMIZATION UNDER MONOPOLY

After describing our model and validate it, we start our analysis with a monopolistic, revenue maximizing service provider, and therefore, drop the subscript j. In the first part of this section, we allow the provider to set both base price and quality level, and in the second part we maximize the provider's revenue for a given base quality level and we provide some numerical examples.

3.1 Optimal Price-Quality Menu

The service provider chooses base quality level q_1, base price level p_1 and number of quality levels to offer L; scaling factor α is endogenous.

The revenue of the monopolistic service provider when she offers only one quality level (q_1):

$$\Pi_1(p_1, q_1) = \sum_{i \in \mathcal{I}(p_1,q_1)} \lambda_i p_1 \frac{w}{q_1},$$

[5]The accuracy of our fit is not dependent on $w = 1000$ assumption. Any w would yield the same accuracy.

where $\mathcal{I}(p_1, q_1)$ is the set of customer types that choose to buy the product when the price is p_1 and the quality level is q_1.

When the service provider offers two quality levels $(q_1, 2\alpha q_1)$, the revenue becomes

$$\Pi_2(p_1, q_1) = \sum_{i \in \mathcal{I}_1(p_1,q_1)} \lambda_i p_1 \frac{w}{q_1} + \sum_{i \in \mathcal{I}_2(p_1,q_1)} \lambda_i p_1 \frac{w}{\alpha q_1},$$

where $\mathcal{I}_1(p_1, q_1)$ is the set of customer types that choose to buy the low quality product and $\mathcal{I}_2(p_1, q_1)$ is the set of customer types that choose to buy the high quality product when the price-quality menu is $\{(p_1, q_1), (2p_1, 2\alpha q_1)\}$.

LEMMA 1. *$\mathcal{I}_1(p_1, q_1) \cup \mathcal{I}_2(p_1, q_1) \supseteq \mathcal{I}(p_1, q_1)$. for any (p_1, q_1).*

PROOF. Suppose $i \in \mathcal{I}(p_1, q_1)$ and $i \notin \mathcal{I}_1(p_1, q_1) \cup \mathcal{I}_2(p_1, q_1)$. If $i \in \mathcal{I}(p_1, q_1)$, then $v_i \geq \frac{c_i + p_1}{q_1}$. If $i \notin \mathcal{I}_1(p_1, q_1) \cup \mathcal{I}_2(p_1, q_1)$ then $v_i < \frac{c_i + p_1}{q_1}$ and $v_i < \frac{c_i + 2p_1}{2\alpha q_1}$. Contradiction. □

PROPOSITION 1. *Offering an additional quality level generates at least as much revenue as offering fewer number of quality levels.*

PROOF. It is enough to show that $\Pi_{k+1}(p_1, q_1) \geq \Pi_k(p_1, q_1)$ for any (p_1, q_1) and $k \geq 1$.

We can easily show the inequality holds for $k = 1$ case, i.e., $\Pi_2(p_1, q_1) \geq \Pi_1(p_1, q_1)$ for any (p_1, q_1).

$$\begin{aligned}\Pi_2(p_1, q_1) &= \sum_{i \in \mathcal{I}_1(p_1,q_1)} \lambda_i p_1 \frac{w_i}{q_1} + \sum_{i \in \mathcal{I}_2(p_1,q_1)} \lambda_i p_1 \frac{w_i}{\alpha q_1} \\ &\geq \sum_{i \in \mathcal{I}_1(p_1,q_1)} \lambda_i p_1 \frac{w_i}{q_1} + \sum_{i \in \mathcal{I}_2(p_1,q_1)} \lambda_i p_1 \frac{w_i}{q_1} \\ &\geq \sum_{i \in \mathcal{I}(p_1,q_1)} \lambda_i p_1 \frac{w_i}{q_1} = \Pi_1(p_1, q_1).\end{aligned}$$

The same procedure follows for any $k > 1$. □

Proposition 1 shows that offering a higher quality level does not cannibalize the service provider's revenue. The next step is to formulate customer preferences on different quality levels.

PROPOSITION 2. *If $c_i \in \left[0, \frac{2p_1(1-\alpha)}{2\alpha - 1}\right)$, then customer type i chooses the first quality level given that her utility is nonnegative. Similarly, if $c_i \in \left[\frac{2^{k-1}p_1(1-\alpha)}{2\alpha-1}, \frac{2^k p_1(1-\alpha)}{2\alpha-1}\right)$, then customer type i chooses quality level k given that her utility is nonnegative.*

PROOF. Let $f_k(c) = \frac{c + 2^{k-1}p_1}{2^{k-1}\alpha^{k-1}}$. $f_k(c)$ is linearly increasing in c for any nonnegative integer k. The slope of $f_k(c)$ is $\frac{1}{2^{k-1}\alpha^{k-1}}$ which is decreasing in k. Hence, for any k, if $\frac{\bar{c} + 2^k p_1}{2^k \alpha^k q_1} \leq \frac{\bar{c} + 2^{k-1}p_1}{2^{k-1}\alpha^{k-1}q_1}$ for a given $\bar{c}$, then the inequality holds $\forall\ c \geq \bar{c}$.

Let c_k be the level such that $f_{k+1}(c_k) = f_k(c_k)$ (which implies $f_{k+1}(c) \leq f_k(c)\ \forall\ c \geq c_k$). If $c_i \in [0, c_1)$, then customer type i chooses the first quality level, and if $c_i \in [c_{k-1}, c_k)$. then customer type i chooses quality level k.

Finally, if $f_{k+1}(c_k) = \dfrac{c_k + 2^k p_1}{2^k \alpha^k} = \dfrac{c_k + 2^{k-1} p_1}{2^{k-1}\alpha^{k-1}} = f_k(c_k)$, then $c_k = \dfrac{2^k p_1 (1-\alpha)}{2\alpha - 1}$ for $0.5 < \alpha_j < 1$. $\square$

So far we have assumed that all customer types have the same valuation v and workload requirement w. From this point on, we make an additional assumption that there is a continuum of customer types that differ with respect to the delay sensitivity parameter, c, where $c \sim U(0, \bar{c})$. Moreover, we assume that the number of quality levels can be at most 4, which is aligned with what we observe in the market.[6]

We start with the revenue maximization problem with one quality level:

$$\begin{array}{ll} \underset{p_1, q_1}{\text{maximize}} & \Pi_1(p_1, q_1) = \dfrac{1}{\bar{c}} \dfrac{p_1}{q_1} \left(\min\{v q_1 - p_1, \bar{c}\}\right) \\ \text{subject to} & 0 < q_1 \le \bar{q_1},\ p_1 \ge 0. \end{array} \tag{1}$$

Let (p_1^*, q_1^*) be the optimal base price and quality level and Π_1^* be the optimal revenue for (1). Then, using first order conditions, it can easily be shown that

$$p_1^* = \begin{cases} \dfrac{v\bar{q}}{2}, & \text{if } \bar{q} \le \dfrac{2\bar{c}}{v} \\ v\bar{q} - \bar{c}, & \text{if } \bar{q} > \dfrac{2\bar{c}}{v} \end{cases},$$

and

$$\Pi_1^* = \begin{cases} \dfrac{v^2 \bar{q}}{4\bar{c}}, & \text{if } \bar{q} \le \dfrac{2\bar{c}}{v} \\ v - \dfrac{\bar{c}}{\bar{q}}, & \text{if } \bar{q} > \dfrac{2\bar{c}}{v} \end{cases}.$$

As the number of quality levels offered increases, the revenue maximization problem gets more complicated, and the closed form solutions have multiple cases. Therefore, we only present the results for the case where there are *exactly* four quality levels assuming $\bar{q}$ is high enough that it is not binding in the problem.[7] The revenue function can be written as

$$\begin{aligned} \Pi_4(p_1, q_1) = \frac{1}{\bar{c}} \frac{p_1}{q_1} \Bigg\{ & \frac{1}{\alpha^3} \left[\min\left\{ 8\alpha^3 v q_1 - 8p_1, \bar{c} \right\} - \min\left\{ \frac{8p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \right] \\ & + \frac{1}{\alpha^2} \left[\min\left\{ \frac{8p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} - \min\left\{ \frac{4p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \right] \\ & + \frac{1}{\alpha} \left[\min\left\{ \frac{4p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} - \min\left\{ \frac{2p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \right] \\ & + \min\left\{ \frac{2p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \Bigg\}. \end{aligned} \tag{2}$$

[6]There are a few product families with 5 or 6 quality levels, such as Google's *n1-standard* and *n1-highmem* families. We only choose 4 quality levels here for simplification.

[7]The formulation presented assumes the third quality level is chosen by at least some customer types.

PROPOSITION 3. *Let the optimal price and base quality level in (2) be (p_1^*, q_1^*). Then*

$$p_1^* = \frac{\bar{c}}{8} \left[\frac{\sqrt{\alpha^6 + 2\alpha^5 - 3\alpha^4 - 8\alpha^2 + 8\alpha}}{\alpha^3 + \alpha^2 - 6\alpha + 4} - 1 \right] \tag{3}$$

and

$$q_1^* = \frac{\bar{c} + 8p_1^*}{8\alpha^3 v}.$$

PROOF. The revenue function can be rewritten as

$$\begin{aligned} \Pi_4(p_1, q_1) = \frac{1}{\bar{c}} \frac{p_1}{q_1} \Bigg\{ & \frac{1}{\alpha^3} \min\left\{ 8\alpha^3 v q_1 - 8p_1, \bar{c} \right\} \\ & - \left(\frac{1}{\alpha^3} - \frac{1}{\alpha^2} \right) \min\left\{ \frac{8p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \\ & - \left(\frac{1}{\alpha^2} - \frac{1}{\alpha} \right) \min\left\{ \frac{4p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \\ & - \left(\frac{1}{\alpha} - 1 \right) \min\left\{ \frac{2p_1(1-\alpha)}{2\alpha - 1}, \bar{c} \right\} \Bigg\}. \end{aligned}$$

First, in the optimal solution, there will be some customer types that choose the highest quality product (by assumption), which implies

$$8\alpha^3 v q_1 - 8p_1 \ge \bar{c} > \frac{8p_1(1-\alpha)}{2\alpha - 1}. \tag{4}$$

Therefore, the revenue function can be written as

$$\begin{aligned} \Pi_4(p_1, q_1) = \frac{1}{\bar{c}} \frac{p_1}{q_1} \Bigg\{ & \frac{1}{\alpha^3} \min\left\{ 8\alpha^3 v q_1 - 8p_1, \bar{c} \right\} \\ & - \left(\frac{1}{\alpha^3} - \frac{1}{\alpha^2} \right) \frac{8p_1(1-\alpha)}{2\alpha - 1} - \left(\frac{1}{\alpha^2} - \frac{1}{\alpha} \right) \frac{4p_1(1-\alpha)}{2\alpha - 1} \\ & - \left(\frac{1}{\alpha} - 1 \right) \frac{2p_1(1-\alpha)}{2\alpha - 1} \Bigg\}. \end{aligned}$$

Assume $(\bar{p}_1, \bar{q}_1)$ is a global maximizer of the function above with $8\alpha^3 v \bar{q}_1 - 8\bar{p}_1 > \bar{c}$. Since $\Pi_4(\bar{p}_1, \bar{q}_1) \ge \Pi_4(0, \bar{q}_1) \ge 0$, it can easily be shown that $\Pi_4(\bar{p}_1, \bar{q}_1 - \varepsilon) \ge \Pi_4(\bar{p}_1, \bar{q}_1)$ for a positive ε. Hence, $8\alpha^3 v q_1^* - 8p_1^* = \bar{c}$ where (p_1^*, q_1^*) is the optimal base price, base quality level couple. Using this equality, the revenue function can be rewritten as

$$\begin{aligned} \Pi_4(p_1) = \frac{1}{\bar{c}} \Bigg\{ & \frac{8\alpha^3 v p_1}{\bar{c} + 8p_1} \left[\frac{1}{\alpha^3} \bar{c} - \frac{8p_1(1-\alpha)^2}{(2\alpha - 1)\alpha^3} \right. \\ & \left. - \frac{4p_1(1-\alpha)^2}{(2\alpha - 1)\alpha^2} - \frac{2p_1(1-\alpha)^2}{(2\alpha - 1)\alpha} \right] \Bigg\}, \end{aligned}$$

where $p_1 \in \left[0, \dfrac{\bar{c}(2\alpha - 1)}{8(1-\alpha)}\right)$. The upper bound is found by using (4). Then,

$$\begin{aligned} p_1^* = \underset{0 \le p_1 < \frac{\bar{c}(2\alpha-1)}{8(1-\alpha)}}{\operatorname{argmax}} \Bigg\{ & \frac{8\alpha^3 v p_1}{\bar{c} + 8p_1} \left[\frac{1}{\alpha^3} \bar{c} - \frac{8p_1(1-\alpha)^2}{(2\alpha - 1)\alpha^3} \right. \\ & \left. - \frac{4p_1(1-\alpha)^2}{(2\alpha - 1)\alpha^2} - \frac{2p_1(1-\alpha)^2}{(2\alpha - 1)\alpha} \right] \Bigg\}, \end{aligned}$$

$\Pi_4(p_1)$ has a global maximum in $\left[0, \frac{\bar{c}(2\alpha-1)}{8(1-\alpha)}\right)$ and the first order conditions give

$$\begin{aligned} p_1^* = \{ & p_1 > 0 \mid \bar{c}^2(2\alpha - 1) = \\ & 4p_1(\bar{c} + 4p_1)(1-\alpha)(4 - 2\alpha - \alpha^2) \}. \end{aligned}$$

The positive root of p_1^* is

$$p_1^* = \frac{-\alpha^3\bar{c} - \alpha^2\bar{c} + 6\alpha\bar{c} - 4\bar{c}}{8(\alpha^3 + \alpha^2 - 6\alpha + 4)} + \frac{\sqrt{\alpha^6\bar{c}^2 + 2\alpha^5\bar{c}^2 - 3\alpha^4\bar{c}^2 - 8\alpha^2\bar{c}^2 + 8\alpha\bar{c}^2}}{8(\alpha^3 + \alpha^2 - 6\alpha + 4)},$$

which is equivalent to (3). □

Proposition 3 shows that as $\bar{c}$ increases, both p_1^* and q_1^* increase. Moreover, as v increases p_1^* does not change while q_1^* decreases. It means that as customers are willing to pay more for the service, instead of increasing the unit price, the provider would deliberately degrade the quality level and sell it with the same unit price, which increases the revenue in return since the processing time becomes longer. The intuition for this result is that increasing the base price makes some customer types choose lower quality levels. Since higher quality products always generate more revenue to the provider, this shift lowers the impact of revenue increase coming from the price increase. On the other hand, decreasing the base quality level does not make any changes on customer preferences and all customer types pays more for the service completion.

Next, we provide sufficient conditions on the optimal number of quality levels to offer under monopoly.

PROPOSITION 4. *Sufficient conditions for offering multiple quality levels:*

i) If $v\bar{q}\left(2\alpha - \frac{2\alpha-1}{\alpha}\right) \le \bar{c}$ *and* $\frac{2}{3} < \alpha < 1$, *offering two quality levels generate more revenue than offering only one quality level.*

ii) If $2\alpha v\bar{q}\left(2\alpha - \frac{2\alpha-1}{\alpha}\right) \le \bar{c}$ *and* $\frac{2}{3} < \alpha < 1$, *offering three quality levels generate more revenue than offering two quality levels.*

iii) If $4\alpha^2 v\bar{q}\left(2\alpha - \frac{2\alpha-1}{\alpha}\right) \le \bar{c}$ *and* $\frac{2}{3} < \alpha < 1$, *offering four quality levels generate more revenue than offering three quality levels.*

PROOF. i) $v\bar{q}\left(2\alpha - \frac{2\alpha-1}{\alpha}\right) \le \bar{c}$ implies $\bar{q} \le \frac{2\bar{c}}{v}$. Hence, the revenue of the one-quality-level case is $\frac{v^2\bar{q}}{4\bar{c}}$. Under the given conditions, the revenue of the two-quality-level case is $\frac{v^2\bar{q}(2\alpha-1)}{2\alpha\bar{c}}$. For positive v and $\bar{c}$,

$$\frac{v^2\bar{q}(2\alpha-1)}{2\alpha\bar{c}} > \frac{v^2\bar{q}}{4\bar{c}}$$

when $\frac{2}{3} < \alpha < 1$.

ii) Let Π_2^* be the optimal revenue when two quality levels are offered, with the optimal base quality level $\bar{q}$ (optimal base quality level has to be equal to $\bar{q}$ when the condition in the first point is satisfied), and the optimal base price p_2^*. Let $\bar{\Pi}_2$ be the revenue when the service provider offers only the second quality level with price $2p_2^*$. Then $\bar{\Pi}_2 > \Pi_2^*$, since the second quality level still gives nonnegative utility to the customer types that chose the first quality level when the first quality level is present, and they pay more than before. Now, assume that the new highest base quality level is $2\alpha\bar{q}$, and follow the first item of this proposition.

iii) The same idea follows.

□

3.2 Optimal Price Menu under Fixed Quality Levels

While controlling both price and quality levels at the same time potentially generates more revenue to the service provider, another interesting question is to find the optimal prices given quality levels. When the service provider offers only one quality level p_1, the optimal price is similar to what we presented in the previous section:

$$p_1^* = \begin{cases} \frac{vq_1}{2}, & \text{if } q_1 \le \frac{2\bar{c}}{v} \\ vq_1 - \bar{c}, & \text{if } q_1 > \frac{2\bar{c}}{v} \end{cases}. \quad (5)$$

When there are two quality levels, $(q_1, 2\alpha q_1)$, the optimal price menu is $(p_1^*, 2p_1^*)$, where

$$p_1^* = \max\left\{\frac{2v\alpha q - \bar{c}}{2}, \frac{vq(2\alpha-1)}{2\alpha}\right\}, \quad (6)$$

only if $p_1^* \le \frac{\bar{c}(2\alpha-1)}{2(1-\alpha)}$; otherwise offering one quality level is preferred to offering two.

When there are more than two quality levels, the optimal price depends on multiple conditions and it is beyond the scope of this exercise. Instead, we provide some numerical examples.

Numerical Examples. In this part, we are going to illustrate cases on how many quality levels the monopolistic service provider offers in the optimal solution given its base quality level, scaling factor, and customer characteristics.[8]

1. If service provider a from the previous section with $(\alpha, q_1) = (0.693, 1.355)$ is the only provider in the market with $v = 0.488$ and $\bar{c} = 0.961$, then the optimal price is indeed \$0.100 and offering 4 quality levels is the revenue maximizing strategy. In other words, if service provider a has $(\alpha, q_1) = (0.693, 1.355)$ and offers 4 quality levels with base price level \$0.100, then, we can infer the market conditions as $v = 0.488$ and $\bar{c} = 0.961$ (using Proposition 3).

2. If service provider b from the previous section with $(\alpha, q_1) = (0.616, 1.733)$ is the only provider in the market with $v = 0.488$ and $\bar{c} = 0.961$ (as above), then the optimal price is \$0.423 and only the base quality level product is being chosen by some customers and the rest choose the *no-buy* option. Setting a price of \$0.126 in this market generates less revenue although all four quality levels are chosen by some customer types and there is no customer type that chooses the *no-buy* option.

[8]The optimal prices found here are searched on a grid with \$0.001 increments. Therefore, the sensitivity of the optimal prices is \$0.001.

These examples show that given market conditions and selected product quality, the monopolistic service provider may choose to offer multiple products (as in Example 1 above) or choose to offer only one product with a price level that may be too high for low customer types (as in Example 2). This behavior is intuitive when the service provider has a relatively high base quality level and a low scaling factor as higher product types do not provide much higher quality than the base quality, which is already high for the market.

4. REVENUE MAXIMIZATION UNDER DUOPOLY

In this section we extend our previous analysis to duopoly case where providers have their own base quality levels and scaling factor set and announced, and they compete with the base price. We still assume that each provider can offer at most 4 quality levels and customers have common valuation v and workload w, and different delay sensitivities $c \sim U(0, \bar{c})$.

We start with a simple model where each provider offers only one quality level. Let (p_1, q_1) and (p_2, q_2) be the price and quality for the first and second providers, respectively. Without loss of generality, assume $q_1 < q_2$. Then, customers with lower type (lower delay sensitivity) choose the first provider, while high types choose the second. Customer type $\hat{c}$ is indifferent between the first and second provider, where

$$\hat{c} = \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1},$$

assuming $\hat{c} \geq 0$.[9] Given p_2, the objective function of the first provider is

$$R_1(p_1) = \frac{1}{\bar{c}} \frac{p_1}{q_1} \hat{c} = \frac{1}{\bar{c}} \frac{p_1}{q_1} \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1}$$

and given p_1, the objective function of the second provider is

$$R_2(p_2) = \frac{1}{\bar{c}} \frac{p_2}{q_2} \left[\max \left\{ \min \left\{ v q_2 - p_2, \bar{c} \right\} - \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1}, 0 \right\} \right].$$

PROPOSITION 5. *Let p_1^e and p_2^e be the equilibrium prices for the first and second provider. Then the Nash equilibrium satisfies*

$$p_1^e = \frac{p_2^e q_1}{2 q_2},$$

and

$$p_2^e = \underset{p_2 \in \{0, p_2^x, p_2^y\}}{\operatorname{argmax}} R(p_2),$$

where $R(p_2)$ is evaluated for $p_1 = p_1^e$, and

$$p_2^x = \max \left\{ v q_2 - \bar{c}, \frac{2 v q_2 (q_2 - q_1)}{4 q_2 - q_1} \right\},$$

$$p_2^y = \min \left\{ v q_2 - \bar{c}, \frac{2 \bar{c} (q_2 - q_1)}{3 q_1} \right\}.$$

PROOF. First, we write down the first order conditions for p_1:

$$p_1^e = \frac{p_2^e q_1}{2 q_2}.$$

Then, we separate the second provider's problem into two cases, $(P1)$ and $(P2)$, and solve both.

$$(P1): \underset{p_2}{\text{maximize}} \quad \frac{1}{\bar{c}} \frac{p_2}{q_2} \left[v q_2 - p_2 - \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1} \right]$$
$$\text{subject to} \quad p_2 \geq v q_2 - \bar{c}.$$

By taking the derivative of the revenue function and then plugging p_1^c for p_1, we reach

$$p_2^x = \max \left\{ v q_2 - \bar{c}, \frac{2 v q_2 (q_2 - q_1)}{4 q_2 - q_1} \right\}.$$

$$(P2): \underset{p_2}{\text{maximize}} \quad \frac{1}{\bar{c}} \frac{p_2}{q_2} \left[\bar{c} - \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1} \right]$$
$$\text{subject to} \quad p_2 \leq v q_2 - \bar{c}.$$

By taking the derivative of the revenue function and then plugging p_1^c for p_1, we reach

$$p_2^y = \min \left\{ v q_2 - \bar{c}, \frac{2 \bar{c} (q_2 - q_1)}{3 q_1} \right\}.$$

If both $(P1)$ and $(P2)$ give a negative revenue in the optimal solution, then, $p_2^e = 0$ which generates zero revenue; otherwise, $p_2^e = \underset{p_2 \in \{p_2^x, p_2^y\}}{\operatorname{argmax}} R(p_2)$ □

As we point out in the previous section, when we allow the service provider to have more than one quality level, the solution depends on v, $\bar{c}$, and the base quality level in a more complicated way. Therefore, it is not straightforward to find closed-form solutions for duopoly case. Instead we simulate the market with different parameters.[10] In our simulation model, first, service provider a from the previous section sets its monopoly price. Second, given a's price, service provider b finds its best response. Then, service provider a finds its best response given b's price, so on and so forth. We iterate this game for 100 times to see if the game reaches a Nash equilibrium that neither of the players would want to change their prices. We analyze four different cases below. In none of the cases we reach a Nash equilibrium. Each case has a different Edgeworth cycle with different price ranges and periodicity. These case are depicted in Figure 2 and described below.

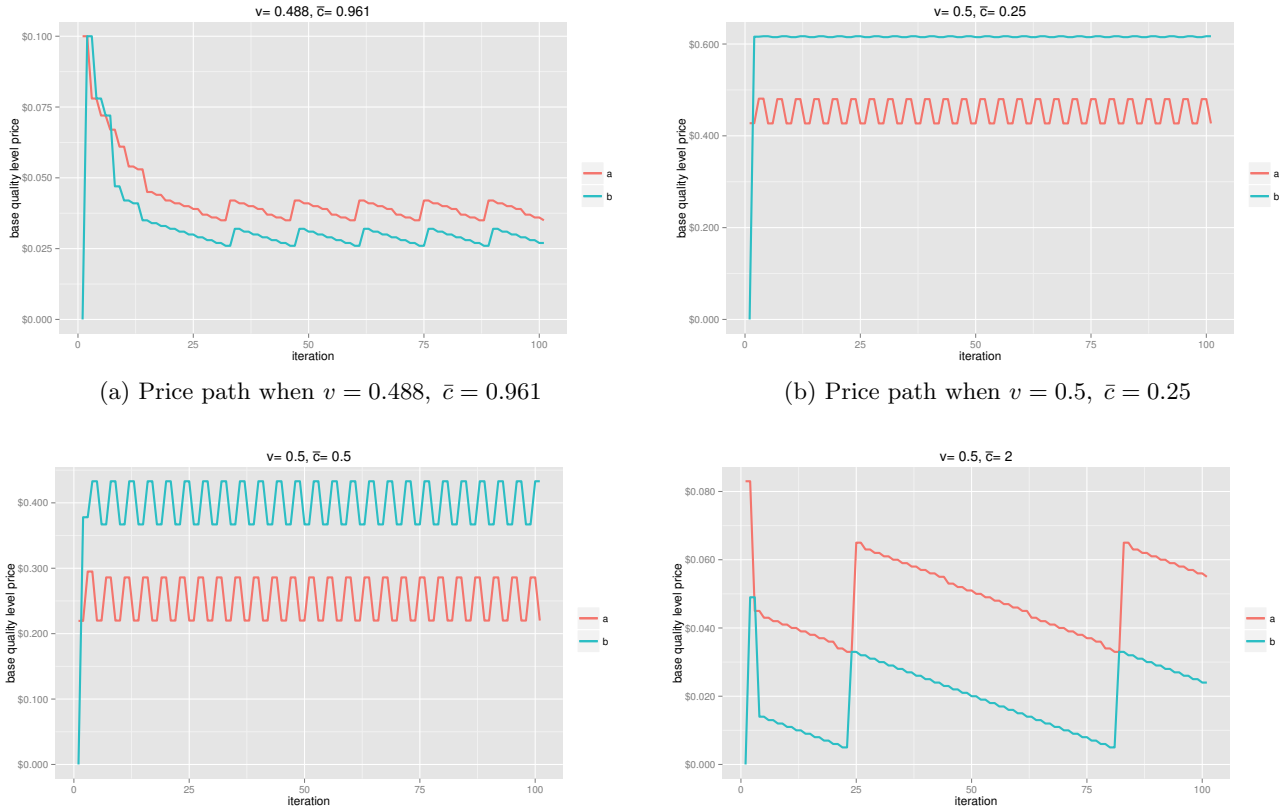
1. $v = 0.488$, $\bar{c} = 0.961$: We have shown that the optimal price for a in this market is \$0.100 when there is monopoly, while it is \$0.423 for b; and we have concluded that if b is the monopoly, there is no point of offering more than one quality level. However, when there is competition, offering more than one quality level becomes preferable to offering only one level for b.

   The price competition makes a decrease its monopoly prices by more than 50%. The price for a varies between \$0.035 and \$0.042 in the cycle, while it is \$0.026 and \$0.032 for b (Figure 2(a)).

2. $v = 0.5$, $\bar{c} = 0.25$: a only uses the base quality level under monopoly, where the optimal price is \$0.4275, which is found using (5). Under duopoly, we have found that only the first quality level is used in both

[9] In Nash equilibrium, $\hat{c}$ is indeed nonnegative, which could be derived using Proposition 5.

[10] As before, we use a price grid with \$0.001 increments.



(a) Price path when $v = 0.488,\ \bar{c} = 0.961$

(b) Price path when $v = 0.5,\ \bar{c} = 0.25$

(c) Price path when $v = 0.5,\ \bar{c} = 0.5$

(d) Price path when $v = 0.5,\ \bar{c} = 2$

Figure 2: Price Paths under Duopoly with Different Parameter Settings

providers in the Edgeworth cycle, and the prices range from \$0.427 to \$0.481 for a, \$0.615 to \$0.617 for b (Figure 2(b)).

Since higher quality levels are not selected in either of the providers, we can assume that each provider offers only one quality level and try to find the equilibrium prices that could potentially be aligned with Figure 2(b). Under this assumption, the equilibrium prices can be calculated by using Proposition 5 as

$$p_1^e = \$0.018 \text{ and } p_2^e = \$0.045.$$

However, when we relax this assumption and let both providers to offer four quality levels, these prices are no longer equilibrium prices because they are so low that high type customers prefer higher quality levels; and therefore, the equilibrium is no longer sustained.

3. $v = 0.5$, $\bar{c} = 0.5$: two quality levels are used in a under monopoly with the optimal price of \$0.220, which is found using (6). Under duopoly, first two quality levels are used in a and only one quality level is used in b in the cycle. The price ranges in the Edgeworth cycle are from \$0.22 to \$0.286 and from \$0.367 to \$0.433 for a and b, respectively (Figure 2(c)).

4. $v = 0.5$, $\bar{c} = 2$: all four quality levels are used in a under monopoly. Under duopoly, all quality levels in both providers are used as well. In this setting, the price varies more for both providers in the Edgeworth cycle (Figure 2(d)).

While the price of a is higher than the price of b in Cases 1 & 4, it is reversed in Cases 2 & 3. It is important to note that the price cycle ranges depend on the initial price level we start the iterative pricing procedure. For instance, if we start Case 2 with a lower price level for provider a, we reach a price cycle with ranges from \$0.005 to \$0.006 and from \$0.003 to \$0.004 for providers a and b, respectively. In this solution, both providers generate lower revenue although all four quality levels are selected by some customer types, which in turn, pushes the prices for provider a to be higher than provider b in the price cycle, contrary to the one quality level case.

5. MODEL EXTENSIONS

There are many avenues to explore by using our price-quality model as a building block. In this paper, we have assumed there is one common workload for all customer types, which implies that the scaling factor, α, only depends on the provider in our model. In reality customers have different workloads and the scaling factor is a combination of the type of workload and the scaling performance of the provider. One potential way to modify the model would be to write the scaling factor as $\alpha = \beta\gamma$, where $\beta \in [0.5, 1]$ is a workload dependent parameter that denotes how parallelizable the workload is, and γ is the scaling factor of the provider.

Assuming that our DaCapo workload has $\beta = 0.8$, since it is moderately parallelizable, γ values become 0.866 and 0.770 for providers a and b, respectively. We have simulated scenarios where β is uniformly distributed between 0.5 and 1 and reached similar results with Edgeworth cycles.

Another extension is to solve profit maximization problem instead of revenue maximization. However, this would add an additional layer of complication on the cost side. At the simplest level, unit cost of a product depends on the configuration that the provider uses, rather than the quality level, and the scale of the provider, since economies of scale plays an important role. With enough information on cost, the model can be modified for profit maximization.

The cloud computing market is a fast growing market and in such market dynamics, sometimes players aim to maximize their market share in the short run before revenue or profit maximization, which could potentially generate higher profits to a player in the long run once the it has its own customer base. In market share maximization case, the duopoly prices are determined based on how much providers can handle profit loss in the short run. In the extreme case, both set prices equal to zero. On the other hand, in zero profit case prices would be set based on costs which may give rise to interesting results as the quality levels and the scaling factors play an important role.

One potential work would be to extend our duopoly game to a two-stage game in which providers first compete in quality and then compete in price.

Fctre3:rkmL(l)11(rg)1reM(l)1c siP1(ri)1(c)1(e)93(Hi(n)1(s)toib)1y(-

pldte416tolrat-

plnctsnrgwotslnceal-
pncton1(:(,586(1(b)-271i)c)1(o)-2d)no-pdts(d)50hran(g(e)5329(t)1(h

6. RECONCILING MODEL PREDICTIONS AND REAL-WORLD BEHAVIOR

As mentioned in the introduction, price cycles in cloud computing are not observed in practice. In the computing, technological advances mean costs are constantly falling. This provides both a market perception that prices should not rise and means that constant prices can be effectively viewed as price increases relative to costs. In Figure 3 we show AWS prices for the "general compute" (M series), large size, with the number indicating the generation. Later generations can only run on newer, higher performance hardware. This new hardware can also run the older generations more cost effectively than before. While this is only one product family, the trends are representative.

A few interesting observations can be derived from the figure. First, in the most recent time period, the best VM sells at the lowest price, whereas the worst sells at the highest price. Second, during the price war period of April 2014, the then-newest generation saw a larger price decrease than the older generation. Finally, the oldest generation is still offered and sold in the marketplace. Relative to the falling prices of new generations, this constant price can be conceptualized as a price increase. While these patterns are certainly not equivalent to Edgeworth cycles, they do evidence "price wars" in one segment of the product space (new generations) and relatively high prices in other parts.

Finally we note some caveats to the realism of our model. We calibrated the model using certain benchmark work-genounen-genotb,

brated model helps not only explain price cutting behavior but also how providers can manage a profit despite predictions that the market "should be" totally commoditized.

8. ACKNOWLEDGMENTS

Any views and opinions expressed herein are solely those of the authors and do not reflect those of the Microsoft Corporation or Columbia University. We thank Omid Alipourfard, Jacob LaRiviere, Jacob Leshno, Hongqiang Liu, Costis Maglaras, and R. Preston McAfee for valuable comments and assistance.

9. REFERENCES

[1] Amazon elastic compute cloud. https://aws.amazon.com/ec2/. Accessed: 2015-09-05.

[2] Cloudharmony. https://cloudharmony.com/. Accessed: 2015-09-05.

[3] Cloudspectator. http://cloudspectator.com/. Accessed: 2015-09-05.

[4] Dacapo benchmarks. http://www.dacapobench.org/. Accessed: 2015-09-05.

[5] Google compute engine. https://cloud.google.com/compute/. Accessed: 2015-09-05.

[6] Microsoft azure virtual machines. https://azure.microsoft.com/en-us/services/virtual-machines/. Accessed: 2015-09-05.

[7] Profitbricks. https://www.profitbricks.com/. Accessed: 2015-09-05.

[8] V. Abhishek, I. A. Kash, and P. Key. Fixed and market pricing for cloud services. *arXiv preprint arXiv:1201.5621*, 2012.

[9] G. M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference*, pages 483–485. ACM, 1967.

[10] J. Anselmi, D. Ardagna, J. Lui, A. Wierman, Y. Xu, and Z. Yang. The economics of the cloud: price competition and congestion. *ACM SIGecom Exchanges*, 13(1):58–63, 2014.

[11] S. M. Blackburn, R. Garner, C. Hoffman, A. M. Khan, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer, M. Hirzel, A. Hosking, M. Jump, H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanović, T. VanDrunen, D. von Dincklage, and B. Wiedermann. The DaCapo benchmarks: Java benchmarking development and analysis. In *OOPSLA '06: Proceedings of the 21st annual ACM SIGPLAN conference on Object-Oriented Programing, Systems, Languages, and Applications*, pages 169–190, New York, NY, USA, Oct. 2006. ACM Press.

[12] A. A. Chien and V. Karamcheti. Moore's law: The first ending and a new beginning. *Computer*, (12):48–53, 2013.

[13] R. J. Deneckere and R. P. McAfee. Damaged goods. *Journal of Economics & Management Strategy*, 5(2):149–174, 1996.

[14] Y. Feng, B. Li, and B. Li. Price competition in an oligopoly market with multiple iaas cloud providers. *Computers, IEEE Transactions on*, 63(1):59–73, 2014.

[15] J. L. Gustafson. Reevaluating amdahl's law. *Communications of the ACM*, 31(5):532–533, 1988.

[16] A. Li, X. Yang, S. Kandula, and M. Zhang. Cloudcmp: shopping for a cloud made easy. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, pages 5–5. USENIX Association, 2010.

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

[18] E. Maskin and J. Tirole. A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica: Journal of the Econometric Society*, pages 571–599, 1988.

[19] K. S. Moorthy. Product and price competition in a duopoly. *Marketing Science*, 7(2):141–168, 1988.

[20] M. D. Noel. Edgeworth price cycles: Evidence from the toronto retail gasoline market*. *The Journal of Industrial Economics*, 55(1):69–92, 2007.

[21] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. The eucalyptus open-source cloud-computing system. In *Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on*, pages 124–131. IEEE, 2009.

[22] Z. Ou, H. Zhuang, J. K. Nurminen, A. Ylä-Jääski, and P. Hui. Exploiting hardware heterogeneity within the same instance type of amazon ec2. In *4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, 2012.

[23] B. P. Rimal, E. Choi, and I. Lumb. A taxonomy and survey of cloud computing systems. In *INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on*, pages 44–51. Ieee, 2009.

[24] J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz. Runtime measurements in the cloud: observing, analyzing, and reducing variance. *Proceedings of the VLDB Endowment*, 3(1-2):460–471, 2010.

[25] A. Shaked and J. Sutton. Relaxing price competition through product differentiation. *The review of economic studies*, pages 3–13, 1982.

[26] G. Wang and T. E. Ng. The impact of virtualization on network performance of amazon ec2 data center. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[27] H. Xu and B. Li. A study of pricing for cloud resources. *ACM SIGMETRICS Performance Evaluation Review*, 40(4):3–12, 2013.