The Effect of Cloud Computing on Marketing of Web-Based Services

Amirreza Fazli Salehi

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Jeffrey D. Shulman, Chair

Amin Sayedi

Oliver J. Rutz

Program Authorized to Offer Degree:

Business Administration

University of Washington

Abstract

The Effect of Cloud Computing on Marketing of Web-Based Services

Amirreza Fazli Salehi

Chair of the Supervisory Committee:

Jeffrey D. Shulman

Department of Marketing and International Business

Cloud computing technology has changed the way web-based firms perform computation tasks. The cloud offers users on-demand computing resources over the internet where they pay only for the resources they use. Cloud computing offers unique features that can have new marketing implications. Two key features of cloud computing are studied in this dissertation: Autoscaling and Spot pricing. Autoscaling allows cloud users the ability to scale their capacity as their demand changes. Using a game theory model, I show autoscaling in cloud computing can intensify competition among entrepreneurs when the likelihood of success for entrepreneurs is high. Alternatively, I find when there is high uncertainty about success in a new market, autoscaling results in higher prices for end users. Spot pricing refers to the discounted selling of unused cloud resources. Spot resources are not guaranteed to remain available for users. The option to diversify resources across multiple pools is offered to users as a way to handle such interruptions. Modeling the effect of interruptions on the cloud providers' pricing decision, I show diversification can be chosen for low likelihood of interruption when users are highly differentiated. The findings of these studies provide clear implications for marketing decisions made by cloud users as well as cloud providers.

Contents

1 Introduction

The National Institute of Standards and Technology defines cloud computing as follows: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" (Mell and Grance, 2011).

More and more companies are adopting the cloud to handle their computational needs. Examples of companies using the cloud include big firms such as Netflix, Airbnb, Pinterest, Samsung, Expedia, and Spotify as well as many small businesses and startups (Gaudin 2015; Bort 2015). In fact, end user spending on cloud services in 2015 was estimated to be above $100 billion (Flood 2013) with an expected annual growth of 44% in workloads (Ray 2013).

Features exclusive to cloud computing services have the potential to change the way firms make their marketing decisions. Cloud computing allows users a variety of options to satisfy their computational needs. On-demand cloud computing allows users on-demand access to computational capacity, such that they only have to purchase what they need. This feature of the cloud uniquely allows users to launch computational capacity at any point they require them.

Consider entrepreneurs deciding whether to launch a web startup in a new market. Getting started requires an investment of time and money into research, development, legal, and other startup expenses prior to knowing whether the company will ever succeed. Furthermore, web and mobile-based startups rely on computational capacity to serve customers. Every interaction a customer has with an application such as a page load, data transfer, and object viewing requires computational resources.

While cloud computing allows startups to outsource their computational costs, startups who ran their tasks in the cloud often needed to decide, and pre-commit to, their computational capacity at the time of purchasing the cloud service. As such, due to the unpredictable nature of startups' demand, the pre-purchased capacity may be excessive or insufficient for the traffic. For example, BeFunky, an online photo editing startup, was featured on a popular social media site three weeks after launch and saw 30,000 visitors in three hours, crashing their servers (Nickelsburg 2016). Such events happen regularly enough that there is a term for it: *the Slashdot effect,* which as Klems,

Nimis, and Tai (2008) describe occurs when a Web startup company is featured on a popular network, resulting in a significant increase in traffic load and causing the startup's servers to slow down or crash. Such a problem can be quite costly as an Aberdeen study found that "a 1-second delay in page load time can result in a 7% loss in conversion and a 16% decrease in customer satisfaction" (Poepsel 2008). Kissmetrics, an analytics company, reports that 1 in 4 people abandon a page if it takes longer than 4 seconds to load (Work 2011). Though capacity can later be increased, the missed demand can be costly. As Amazon CEO Jeff Bezos describes, startups face a serious challenge when choosing computational capacity:[1]

> "And you do face this issue (demand uncertainty) whenever you have a startup company. You want to be prepared for lightning to strike because if you're not, that generates a big regret. If lightning strikes and you weren't ready for it, that's kind of hard to live with. At the same time, you don't want to prepare your physical infrastructure to hubris levels either in the case that lightning doesn't strike."

To address this challenge, cloud providers such as Amazon, Microsoft, and Google have begun to offer *autoscaling*, a feature that allows startups to scale their computational capacity up or down automatically in real time. Using autoscaling, startups can maintain application availability and scale their computational capacity for serving consumers without having to make capacity pre-commitments.

For companies, autoscaling means having just the right number of servers required for meeting the demand at any point in time, which can provide an attractive solution to handling uncertain demand. Autoscaling is offered with no additional fees and has been celebrated as one of the most beneficial features of cloud computing. Users of autoscaling such as AdRoll and Netflix find that when a new customer comes on board, they can handle the additional traffic instantly.[2] As Mikko Peltola, the Operations Lead at Rovio, noted regarding the benefits of autoscaling, "We can scale up as the number of players go up ... so we can automatically increase the processing power for our servers."[3] The Chief Technology Officer for Cloud at General Electric has mentioned this feature as one of the main reasons for the company's move to the cloud, stating "Running inside a public cloud

[1] https://animoto.com/blog/news/company/amazon-com-ceo-jeff-bezos-on-animoto/, accessed September 2016.
[2] See https://aws.amazon.com/solutions/case-studies/adroll/
[3] See https://aws.amazon.com/solutions/case-studies/rovio/

environment, you're able to consume unlimited capacity as needed" (Weaver 2015). Anecdotally, some entrepreneurs such as Animoto CEO, Brad Jefferson, who used Amazon Web Services (AWS), see the scaling it offers as a game changer for their startup: "We simply could not have launched Animoto.com and our professional video rendering platform at our current scale without massive CapEx and a lot of VC funding. The viral spike in Animoto video creations we experienced this week would have been disastrous without AWS."[4]

Another key feature of the cloud is spot resources. While on-demand cloud computing resources stay active as long as the user requires them, cloud services also provide another cheaper offer in the form of spot resources. Spot resources are unused cloud resources that can be purchased on the spot market for a discount, but are not guaranteed to last throughout the duration that the user requires them. This feature of the cloud allows users to take into account their potential losses from having their resources interrupted when purchasing from the cloud and weight them against their potential savings from discounted spot resources. Furthermore, cloud providers uniquely offer the option of using spot resources from different pools of servers in order to reduce the probability that all of one user's resources get interrupted in case a particular pool of resources becomes unavailable.

Consider a firm using the spot market to purchase low price computing resources. These resources can become interrupted for a multitude of reasons. For instance, there could be outages in the cloud provider's servers.[5] Or, a pool of resources might be requested from a buyer of on-demand services, which would be prioritized over spot users. The result of such interruptions can be harmful to spot users, especially if most or all of their resources get shut down.[6] The diversification feature in cloud computing allows the user to spread their resources across multiple pools, lowering the number of the user's resources that would be interrupted as each pool becomes unavailable.

Each of these unique features offered by cloud computing providers can have new marketing implications for firms using the cloud. Autoscaling allows startups with uncertain demand to receive just the right amount of capacity depending on what their demand turns out to be, removing the need for pre-commitment to capacity. Spot resources allow users to diversify their resources over a variety of pools in order to manage the probabilities of resources becoming unavailable.

In this dissertation, I study the marketing implications of each of the two aforementioned

[4]https://animoto.com/blog/news/company/amazon-com-ceo-jeff-bezos-on-animoto/
[5]https://searchcloudcomputing.techtarget.com/feature/Cloud-computing-outages-What-can-we-learn
[6]https://www.zdnet.com/article/cloud-computing-heres-how-much-a-huge-outage-could-cost-you/

features of cloud computing. In Essay 1, I study autoscaling and how it affects startups' decisions to enter new markets. In Essay 2, I analyze spot pricing and the users' decisions to diversify cloud resources across multiple pools. Finally, I conclude with providing directions for future marketing research on cloud computing.

The findings from this study have implications for various players in the cloud computing market. Users of cloud services can use the findings to decide what type of cloud solution, if any, would minimize their cost of computation. My findings also help providers decide when it is profitable to offer various features to the users. Cloud computing has emerged as a fast growing industry, expected to play a big role in the future of computation. This allows for many research opportunities on the adoption of the cloud. However, so far few papers in the field of marketing have looked at this issue. I believe this study can open the door for more business and marketing research in the future on cloud computing.

2 Literature Review

Academic research on cloud computing is still relatively new and most of the work done on this topic focuses on technological issues of the cloud (e.g., Yang and Tate 2012). The few existing business and economics studies of cloud computing have mainly offered conceptual theories and evidence from surveys and specific cases (e.g., Leavitt 2009; Walker 2009; Gupta, Seetharaman, and Raj 2013). Sultan (2011) suggests that cloud computing can benefit small companies due to its flexible cost structure and scalability. Regarding the ability to autoscale, Armbrust et al. (2009) suggest that elasticity in the cloud shifts the risk of misestimating the workload from the user to the cloud provider. Regarding market entry, Marston, Li, Bandyopadhyay, Zhang, and Ghalsasi (2011) conceptually argue that cloud computing can reduce costs of entry and decrease time to market by eliminating upfront investments. This dissertation is the first to model autoscaling and spot pricing in the cloud and study their marketing implications.

In addition to cloud computing, this research is related to a number of topics in the literature. In particular, previous research shows uncertainty in demand plays an important role in capacity and production decisions. Che, Narasimhan, and Padmanabhan (2010) find how demand uncertainty and consumer heterogeneity affect whether a firm optimally adopts a make-to-stock

system, a backorder system, or a combination of both. Ferguson and Koenigsberg (2007) examine how a firm should sell its deteriorating perishable inventory and compare this option to discarding the previously unsold stock. Desai, Koenigsberg, and Purohit (2007) find the optimal inventory with demand uncertainty as a function of a product's durability. Desai, Koenigsberg, and Purohit (2010) find a strategic reason for retailers to carry inventory larger than the expected sales in both high and low demand states. Biyalogorsky and Koenigsberg (2014) consider product introductions by a monopolist facing uncertainty about consumer valuations and find whether the firm offers multiple products simultaneously or sequentially. Anupindi and Jiang (2008) consider capacity investments by competing firms with demand uncertainty and look at the effect of flexibility to postpone production until demand is realized. They find the flexibility increases capacity investment and profitability, whereas we find when autoscaling may decrease equilibrium computational expenditures and when it may decrease profitability. Relative to these models, autoscaling affects more than the level of uncertainty in demand at the time capacity is set, it eliminates the capacity decision entirely and converts the capacity costs from being sunk at the pricing decision to being variable costs. Moreover, autoscaling does not eliminate uncertainty about firms' success at the entry stage.

The examination of how cloud computing with autoscaling affects a startup's entry decision also relates to the literature on market entry. A body of literature looks at the timing of entry and how an incumbent can deter entry (e.g., Spence 1977; Narasimhan and Zhang 2000; Joshi, Reibstein, and Zhang 2009; Milgrom and Roberts 1982; and Ofek and Turut 2013). In contrast, our model examines simultaneous entry decisions by startups. Our model adds to entry literature by jointly considering both entry and capacity decisions, such that each firm's decision to enter depends on the expected future capacity of both firms and whether this capacity will be chosen ex ante or autoscaled to demand.

I compare autoscaling with cases where firms commit to their computational capacity before pricing and realizing demand. This relates to other papers examining capacity commitments. Kreps and Scheinkman (1983) find that a Bertrand pricing game is equivalent to a Cournot game when capacity is chosen prior to pricing. Reynolds and Wilson (2000) extend this model to include uncertainty about market size and find there is no symmetric, pure-strategy equilibrium capacity choice when there is significant demand variation. Nasser and Turcic (2015) find symmetric horizontally

differentiated firms use asymmetric strategies on whether to commit to capacity or not. This is consistent with our finding in Lemma 2. However, since they do not allow for demand uncertainty, capacity commitments always alleviate competition in their model, whereas, in ours, capacity commitments sometimes intensify competition. Furthermore, at least one firm commits to capacity in any equilibrium in their model, whereas, in our model, both firms may use autoscaling. In asymmetric games, Daughety (1990) finds the Stackelberg leader will commit to a greater quantity, whereas Shulman (2014) finds an unauthorized seller procuring diverted units from authorized retailers will unilaterally choose a lesser quantity than it could profitably sell. Swinney, Cachon, and Netessine (2011) examine the optimal timing of capacity investment in a model in which market price is given by a demand curve and firms can choose to set capacity early at one marginal cost of capacity or after demand is realized at a different cost of capacity. Van Mieghem and Dada (1999) allow firms to choose the time of their pricing decisions and find that postponing pricing until after demand is realized makes the capacity decision less sensitive to demand uncertainty. Our research expands this literature by considering demand uncertainty and market entry decisions in a horizontally differentiated market with and without capacity commitments. In contrast to the previous literature where capacity commitments always alleviate competition, I show that, depending on the level of demand uncertainty, capacity commitments may indeed intensify the competition. Furthermore, in this model, firms decide whether to adopt autoscaling or to pre-commit to capacity; I find that firms do not always follow symmetric strategies in regards to adoption of autoscaling. Finally, I uniquely explore how entry decisions are affected by existence of capacity commitments under demand uncertainty.

Autoscaling in cloud computing also has the effect of converting up-front capacity costs to variable costs that change with demand. Prior research examining the effect of converting fixed to variable costs via outsourcing (e.g., Shy and Stenbacka 2003; Buehler and Haucap 2006; Chen and Wu 2013) have found that prices and profitability rise with this conversion. Relative to these models, demand is uncertain in our model and our up-front capacity cost is endogenous since firms can choose their capacity. We also find, in contrast, that average prices can fall when cloud computing with autoscaling is used even in conditions for which entry is unaffected.

A model of cloud adoption can be related to outsourcing models in the literature. Grossman and Helpman (2002) develop a model where producers can choose between vertical integration

and outsourcing to a provider. In their model, specialized firms which outsource to suppliers can produce goods with lower marginal costs, but incur a costly search for a provider. They show that when firms are identical, in equilibrium no industry has both vertically differentiated firms and firms that use outsourcing. Van Mieghem (1999) considers the capacity decisions of the user firm and the provider, and finds contracts that coordinate their capacity investment decisions in a single-period model. Feng and Lu (2011) show higher cost efficiency of the provider can result in lower profit for the user firms using outsourcing, mainly due to their weakened bargaining power. Wu and Zhang (2014) compare outsourcing to efficient providers, where costs are lower, with outsourcing to responsive providers, where demand information is more accurate. They show responsive outsourcing becomes more attractive relative to efficient outsourcing when market size shrinks or costs of outsourcing increase. Previous literature in economics and operation research has studied contract design in outsourcing and supply chain management (Cachon and Netessine 2003). Hasija, Pinker, and Shumsky (2008) study the different outsourcing contracts for call centers, including pay-per-call and pay-per-time contracts, in the existence of information asymmetry on productivity. Ren and Zhou (2006) consider contracts used by user firms to make the provider use staffing and effort levels optimal for the outsourcing supply chain. They show that while a pay-per-use contract can result in optimal staffing, it cannot induce optimal service quality. Cachon and Harker (2002) study outsourcing contracts in a queuing model, in which the user firm guarantees minimum demand to the provider and the provider guarantees minimum service level. Partial outsourcing, or co-sourcing, refers to the practice of outsourcing some production components and keeping the rest of the components in-house (Aksin, Armony, and Mehrotra 2007). Aksin, de Vericourt, and Karaesmen (2006) study partial outsourcing for a call center and compare contracts which reserve a specific capacity in the provider with contracts which only outsource the overflow calls. While Aksin, de Vericourt, and Karaesmen (2006) assume all customer calls have the same value, Gans and Zhou (2007) consider partial outsourcing in the context of a call center serving their high-value customers in-house and outsourcing the service to its low-value customers. Shy and Stenbacka (2005) consider competition among user firms and show that intensified competition on the users' final good can enlarge the set of outsourced components and result in fewer components produced in-house.

The model of spot pricing includes uncertainties in the availability of cloud resources. This

concept is related to the existence of random yields from suppliers. Yano and Lee (1995) provide a summary of the literature on determining procurement quantities under random yields. Parlar and Berkin (1991) model outages in supply by assuming an interval of random length, during which supply is unavailable and derive the optimal order quantity. One way for retailers to deal with supply uncertainty is diversification and using multiple supply sources (Anupindi and Akella 1993). Babich, Burnetas, and Ritchken (2007) show that in a model of two competing suppliers, when interruptions in the two suppliers are highly correlated the supply price lowers due to intensified competition between suppliers. Dada, Petruzzi, and Schwarz (2007) compare reliable and unreliable suppliers and show that quantity ordered is higher when suppliers are unreliable. Federgruen and Yang (2009) find the optimal choice of unreliable suppliers when the objective function of the procurer is to have a minimum number of resources with a certain probability. Feng and Shi (2012) compare the effectiveness of supply diversification with retailer dynamic pricing as responses to supply uncertainty. He, Huang, and Yuan (2016) model competition among buyers in the presence of supply uncertainty. Prior literature on supply diversification focuses on diversifying across multiple supply sources, whereas in the context of cloud computing diversifying occurs within one supply source. Moreover, cloud computing allows for dynamic pricing from suppliers such that prices can change as interruptions occur, whereas in previous studies suppliers are constrained by contract to a fixed price that does not change as interruptions occur.

3 Essay 1: The Effects of Autoscaling in Cloud Computing on Entrepreneurship

3.1 Introduction

Technology startups often rely on computational resources to serve their customers, though rarely is the number of customers they will serve known at the time of market entry. Today, many of these computations are run using cloud computing. A recent innovation in cloud computing known as autoscaling allows companies to automatically scale their computational load up or down as needed. We build a game theory model to examine how autoscaling will affect entrepreneurs' decisions to enter a new market and the resulting equilibrium prices, profitability, and consumer surplus. Prior to autoscaling, startups needed to set their computational capacity before realizing their computational demands. With autoscaling, a company can be assured of meeting demand and pay only for the demand that is realized.

In this paper, we develop an analytical model to examine how the emergence of autoscaling in cloud computing will affect entrepreneurs' decisions regarding market entry and prices. In particular, despite popular belief, we identify conditions for when autoscaling negatively affects entrepreneurship. In other words, fewer startups will enter in certain markets due to the advent of autoscaling. Autoscaling has several properties that make it unique from some previously explored areas in marketing and operations. In particular, autoscaling:

- removes a capacity decision that otherwise has to be made prior to pricing;
- converts computational capacity costs from fixed costs to variable costs at the time of pricing;
- allows capacity to be set after the uncertainties regarding consumers' level of interest and competitors' strategies are resolved; however, autoscaling still
- preserves the uncertainty that exists at the time of making the entry decision.

Given these properties, the effects of autoscaling on company strategies cannot be addressed by prior research on capacity choices and demand uncertainty. In fact, *our model and predictions diverge from prior literature substantively.*

We uniquely incorporate these properties of autoscaling into a game theory model in which two horizontally differentiated startups have the option to enter a market upon incurring an entry cost. We compare a model in which startups choose computational capacity to a model in which startups can choose to adopt autoscaling in cloud computing. The model is constructed to address the following research questions:

1. How does autoscaling affect entrepreneurs' profits?

2. How does autoscaling affect entrepreneurs' pricing strategies?

3. How does autoscaling affect market entry decisions?

4. How does autoscaling affect consumer surplus?

We explore the roles of several important market factors in determining the answer to each of the research questions. First, we model the startups' ex ante likelihood of a successful venture. As Griffith (2014) suggests, the value that a startup brings to the market is unknown before entry. In some markets, consumer needs are well known and established, therefore yielding a higher likelihood of successfully creating a product that matches consumer needs. For other markets, the consumer needs are less understood and there is a lower likelihood of a successful venture. We show that the likelihood of a successful venture plays a critical role in determining how autoscaling affects equilibrium strategies and profits.

Secondly, we model the degree of potential success a venture may enjoy. For some markets, the consumer need addressed by the startup may have greater value to the customer than other markets. A greater value to the customer can translate to a highly successful venture in terms of the ability to charge higher prices. We examine how the magnitude of a successful venture (in terms of how much a consumer would pay to satisfy the need if the startup successfully addressed it) will affect pricing decisions, market entry, and consumer surplus.

In answering the first research question, we find that autoscaling can increase or even decrease entrepreneurs' expected profits. We identify strategic consequences of autoscaling and the conditions that lead to each possibility. In particular, when the probability of success is sufficiently low, autoscaling increases the entrepreneurs' expected profits. However, autoscaling may also create a prisoner's dilemma situation where startups choose autoscaling, but autoscaling lowers their equi-

librium profits. In particular, when entry costs are sufficiently small and the probability of success is moderately high, startups adopt autoscaling in equilibrium; however, their equilibrium profits would be higher if autoscaling was not available.

In addressing the second research question, we find that autoscaling may increase or decrease average prices charged by competing startups. Existing economic theory would suggest that removing the capacity decision prior to pricing would result in a shift from a Cournot game to a Bertrand game, thereby decreasing prices (e.g., Kreps and Scheinkman 1983). However, our model shows that when entry costs are low enough such that both startups enter the market, autoscaling increases the average prices set by each startup if the probability of a successful venture is not too high. On the other hand, if the probability of a successful venture is sufficiently high, then autoscaling decreases the average prices set by each startup.

With regard to the third research question, we find that autoscaling can sometimes decrease market entry. Though we confirm common intuition that the likelihood of a market being served by at least one startup is improved with autoscaling, we find that, under certain conditions, entry by multiple startups will not occur because of autoscaling. The counter-intuitive result occurs when entry costs are moderately high and there is a high probability that entrants will have a successful venture.

Finally, in addressing the fourth research question, we show that autoscaling may increase or decrease expected consumer surplus depending on the likelihood of a successful venture and entry costs. Given the fact that autoscaling guarantees companies have the capacity to serve consumers in case of high demand, thereby resolving issues such as the Slashdot effect, one might expect that consumers benefit from autoscaling. However, our model shows that when entry costs are low enough such that both startups enter the market, autoscaling decreases expected consumer surplus if and only if the probability of a successful venture is moderate.

The findings of this study provide implications for various players in new markets, including startups, cloud providers and consumers. Our analysis informs startup managers on how autoscaling affects competitive dynamics in pricing and entry. The results suggest that a startup manager should consider not only the positive direct effect of autoscaling in reducing costs, but also the negative strategic effect in altering the nature of competition. By evaluating the probability of success, the value to the consumers, and entry costs, managers can use the findings from this study

to determine whether autoscaling increases or decreases the likelihood of monopoly power over the new market. Our findings also inform cloud providers about how autoscaling affects not only the number of firms using the cloud, but also the number of servers each of those firms purchases. Our model provides insights for consumers on how autoscaling affects the prices charged in the market, showing that for high probabilities of success average prices decrease with autoscaling and for lower probabilities of success they increase with autoscaling. We also find conditions for which autoscaling will decrease or increase consumer surplus, which can be used for consumer surplus maximizing policy design. To the best of our knowledge, this paper is the first to study the marketing aspects of cloud computing, and how it can affect prices and market entry. With the growing trend of adopting the cloud by firms, cloud computing is expected to become a major part of any business and this provides the field of marketing with a variety of related new topics to explore.

In addition to contributions to practice, our work contributes to economic theory regarding capacity commitments. Our benchmark model uniquely solves a capacity choice game with demand uncertainty and horizontal differentiation between sellers. Contrary to the previous literature where capacity commitments lead to higher prices, we show that under demand uncertainty, capacity commitments (relative to autoscaling) can intensify the competition and cause lower equilibrium prices. We also uniquely study the effects of removing capacity commitments made under demand uncertainty (via autoscaling) on firms' market entry decisions.

The rest of this essay is organized in the following order. In Section 3.2, we introduce the model. In Section 3.3, we present the analysis of the model and derive the results. Finally, the discussion of our findings is presented in Section 3.4.

3.2 Model

We consider two symmetric startups who could potentially enter a particular web or mobile application market. Following the convention and for ease of exposition, we subsequently refer to these startups as firms. To enter the market, the firms would incur a fixed entry cost, F. This cost includes start-up expenses such as legal, research and development, and human capital investments. To model post-entry competition, we adopt a discrete horizontal differentiation model (e.g., Narasimhan 1988; Iyer, Soberman, and Villas-Boas 2005; Zhang and Katona 2012; Zhou, Mela, and Amaldoss 2015) with three consumer segments, each consumer demanding at most one unit of

the product. Upon entry, each Firm i will find a segment of consumers, Segment i with $i \in \{1, 2\}$, who will buy from Firm i if and only if the price p_i is below their reservation value v_i and who will derive zero value from the competitor's product. This captures the reality that consumers vary in their taste preferences regardless of firm entry and that firms have idiosyncratic differences that will allow them to serve these tastes differently from each other upon successful entry. The size of each Segment i is given by $\alpha < 1/2$ for $i \in \{1, 2\}$. The remaining $1 - 2\alpha$ consumers, Segment 3, are indifferent between firms and will prefer to buy from the firm with the lowest price. The parameter α can be interpreted as the extent to which consumers vary in their taste preferences.

In the absence of autoscaling, the timing of the game is as follows:

Stage 1: Firms each make a simultaneous observable decision of whether or not to enter the market and thereby incur the cost F. We allow for uncertainty in whether a firm will find the venture successful in terms of whether v_i is high or low. We assume the ex ante probability of a firm finding success in this market is γ, which is common knowledge. In other words, if Firm i enters the market, v_i is an i.i.d. draw from a binary distribution in which $v_i = V$ with probability γ and $v_i = 0$ with probability $1 - \gamma$. This assumption reflects the idea that the value provided to customers is unclear for potential entrants. As Lilien and Yoon (1990) argue, the fit between market requirements and the offering of the new entrant is highly unpredictable and is critical to the success of the entrant. In a survey of 101 startups, it was reported that the number one reason for the failure of a startup is the lack of market need for the offered product (Griffith 2014), suggesting that the value created for customers is unknown to many startups before entry.

To remark on the structure of demand and uncertainty, notice that our model set up has several desirable properties. In particular, it allows a firm to be uncertain about the size of the potential market and the effect of its price on realized demand: the firm may find itself a monopolist, the firm may find itself with very low demand (normalized to zero), or the firm may find itself competing head-to-head. Though one can explore alternative model specifications to capture these same properties, the current specification allows for tractability while uncovering a novel mechanism.

Stage 2: Firms that enter make a simultaneous observable decision of computational capacity. Firms incur a computational cost, c, for each consumer they serve and choose their computational capacity k_i to maximize expected profit.

Stage 3: The reservation value for each Firm i, v_i, becomes common knowledge and each firm in

the market simultaneously chooses p_i to maximize profit.

Stage 4: Demand is realized. In the case a firm experiences demand greater than its computational capacity, we assume an efficient rationing rule (see Tirole 1988, p. 213) in which demand from Segments 1 and 2 is satisfied prior to demand from Segment 3. Residual demand from Segment 3 is allocated to the competing firm, provided it has available capacity.

When autoscaling is possible, we allow firms to decide whether to use autoscaling or choose a computational capacity in Stage 2. If a firm chooses autoscaling, it incurs the computational cost c only on each unit of realized demand. In practice, changing capacity decisions in the absence of autoscaling takes at least a few hours, and in some cases days, before coming into effect on the cloud servers. The Befunky example, the Slashdot effect, and the "lightning strike" analogy by Amazon's CEO, discussed in the introduction, highlight the fact that demand often changes faster than what firms can respond to in terms of computational capacity. Our assumption that capacity decision is made before the demand is realized captures this reality. However, our main results are robust to this assumption. In particular, even if firms can choose to adopt autoscaling after the demand is realized, our results in Propositions 2, 3, 4 and 5 still hold.

The timing of the game is depicted in Figures 1 and 2. A summary of notation is in Table 1.

Symbol	**Description**
α	Size of each of Segments 1 and 2
k_i	Capacity chosen by Firm i
v_i	Reservation value consumers have for Firm i
V	The consumer reservation value in the high-value condition
γ	Probability that $v_i = V$
c	Cost per unit of computational capacity
p_i	Price chosen by Firm i
F	Cost of entry

Table 1: Summary of notation

3.3 Analysis

Our research objective is to identify how the advent of autoscaling affects equilibrium prices, profits, and market entry. To this end, we first examine equilibrium entry and prices in the situation in which computational capacity must be determined prior to demand realization. We will subsequently characterize the equilibrium when autoscaling is available. We will conclude with a

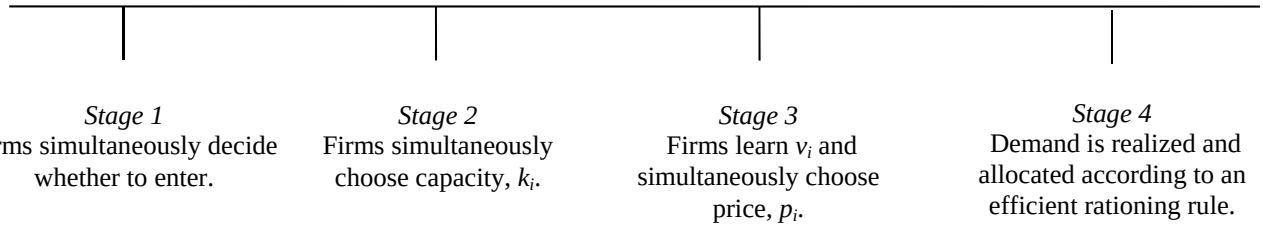

Figure 1: Sequence of events with no autoscaling

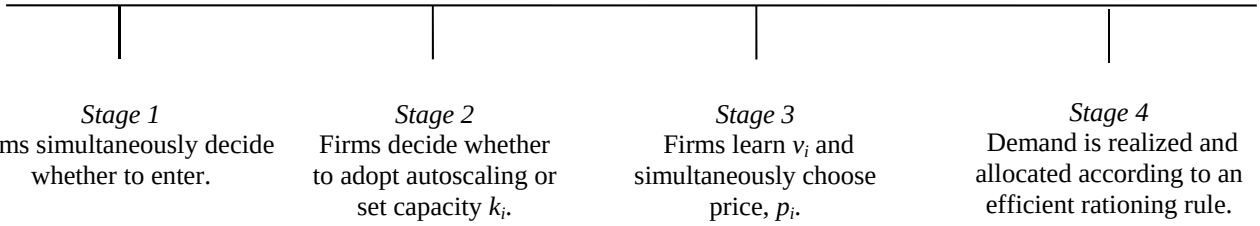

Figure 2: Sequence of events with autoscaling

comparison across these possibilities.

3.3.1 No Autoscaling

We solve the model in which there is no autoscaling via backward induction beginning with the pricing subgame equilibrium. First suppose that $k_1 + k_2 > 1$. We want to calculate equilibrium prices of this game. Without loss of generality, assume that $k_2 \geq k_1$. Also, it is easy to see that firms never set their capacity to $k_i > 1 - \alpha$ or $k_i < \alpha$; therefore, it is sufficient to consider the case where $k_i \in [\alpha, 1 - \alpha]$ for $i \in \{1, 2\}$.

We start by showing that this game does not have a pure strategy equilibrium. Assume for sake of contradiction that the firms use prices p_1 and p_2 in a pure strategy equilibrium. If $p_1 \neq p_2$, then the firm with a lower price can benefit from deviating by increasing its price to $\frac{p_1+p_2}{2}$. If $p_1 = p_2$, then Firm 2 can benefit from deviating by decreasing its price to $p_2 - \varepsilon$, for sufficiently small ε, to acquire more consumers from Segment 3. Therefore, a pure strategy equilibrium cannot exist.

Next, we find a mixed strategy equilibrium for this game. Mixed strategies can be interpreted as sales or promotions and are common in the marketing literature (e.g., Chen and Iyer 2002, Iyer, Soberman, and Villas-Boas 2005, Zhang and Katona 2012). Provided $k_1 \leq 1 - \alpha$, Firm 2 can choose

to *attack* with a price that clears its capacity or *retreat* with a price V that harvests the value from the $1-k_1$ consumers that Firm 1 cannot serve due to its capacity constraint. Let z be the price at which Firm 2 is indifferent between *attacking* to sell to k_2 consumers at price z and *retreating* to sell to $1-k_1$ consumers at price V. We have $z = \frac{V(1-k_1)}{k_2}$. Figures 3 and 4 demonstrate the different appeals of these two pricing strategies. The choice between *retreating* and *attacking* for each firm depends on the choice of the other firm. If Firm 1's price is high, it becomes easier for Firm 2 to attract the consumer segment that is in both firms' reach resulting in Firm 2 choosing to *attack*. On the other hand, if Firm 1's price is low, Firm 2 would prefer to *retreat* than to compete with Firm 1 over the overlapping consumers. In the equilibrium that we find, both firms use a mixed strategy with prices ranging from z to V. Suppose that $F_i(.)$ is the cumulative distribution function of price set by Firm i and $F_j(.)$ is the cumulative distribution function of price set by competing firm j. The profit of Firm i earned by setting price x, excluding the sunk cost of capacity, is

$$\pi_i(x) = F_j(x)(1-k_j)x + (1-F_j(x))k_i x.$$

Using equilibrium conditions, we know that the derivative of this function must be zero for $x \in (z, V)$. Therefore, we have

$$-x(k_i+k_j-1)F_j'(x) - F_j(x)(k_i+k_j-1) + k_i = 0.$$

The solution to this differential equation is

$$F_j(x) = \frac{k_i}{k_i+k_j-1} + \frac{C_j}{x}$$

where constant C_j is determined by the boundary conditions. As for the boundary conditions, we use $F_1(V) = 1$. Therefore, we get

$$F_1(x) = \begin{cases} 0 & \text{if } x < z \\ \frac{(k_1-1)V + k_2 x}{x(k_1+k_2-1)} & \text{if } z \le x < V \\ 1 & \text{if } x \ge V \end{cases} \tag{1}$$

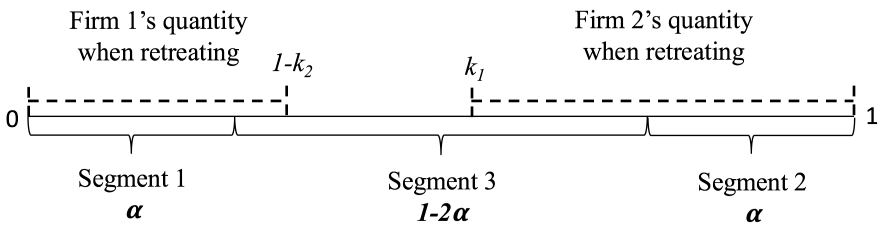

Figure 3: How k_1 and k_2 affect firm incentives to adopt a *retreating* price

This implies that Firm 1 mixes on prices between z and V such that Firm 2 is indifferent between using any two prices in this range. Furthermore, given $F_1(.)$, Firm 2 strictly prefers any price in $[z, V]$ to any price outside this interval. To have an equilibrium, the strategy of Firm 2 should be such that Firm 1's strategy is not suboptimal. In other words, Firm 1 should be indifferent between any two prices in $[z, V]$, and should weakly prefer any price in $[z, V]$ to any price outside this interval. Therefore, we have to use boundary condition $F_2(z) = 0$ to make sure that (1) Firm 1's indifference condition is satisfied in $[z, V]$, and (2) Firm 2 does not set the price to lower than z, as we already know from $F_1(.)$ that such prices are suboptimal for Firm 2. As such, we get

$$F_2(x) = \begin{cases} 0 & \text{if } x < z \\ \frac{k_1((k_1-1)V+k_2x)}{k_2x(k_1+k_2-1)} & \text{if } z \leq x < V \\ 1 & \text{if } x \geq V \end{cases} \tag{2}$$

Note that $F_2(x)$ is discontinuous at $x = V$, and jumps from $\frac{k_1}{k_2}$ to 1. This implies that Firm 2 uses price V with probability $1 - \frac{k_1}{k_2}$. In other words, $f_2(V) = (1 - \frac{k_1}{k_2})\delta(0)$, where $f_2(.)$ is the probability density function for price of Firm 2 and $\delta(.)$ is Dirac delta function.[7] [8]

Given $F_i(.)$, we can calculate the expected profit of each firm in this mixed strategy equilibrium. Excluding the sunk cost of capacity, we have

$$\pi_1 = \frac{(1-k_1)k_1V}{k_2} \quad \text{and} \quad \pi_2 = V(1-k_1).$$

[7] See Hassani (2009), pp 139-170.

[8] One might wonder if the probability density $1 - \frac{k_1}{k_2}$ allocated to price V by Firm 2 could be instead allocated to price z. The answer is that it cannot. While such strategy would still keep Firm 1 indifferent between any two prices in $[z, V]$, it would make price $z - \varepsilon$ (for sufficiently small ε) a strictly better strategy for Firm 1, which violates equilibrium conditions.

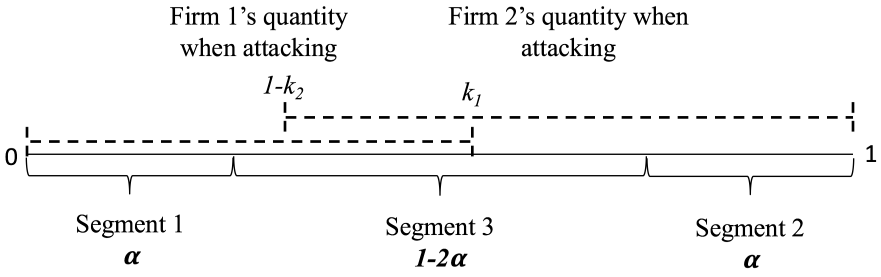

Figure 4: How k_1 and k_2 affect firm incentives to adopt an *attacking* price

Note that the higher capacity firm, Firm 2, earns a profit equal to the profit it would have made if it had chosen a *retreating* strategy while pricing at V; as such, the expected profit of Firm 2 is independent of its capacity k_2. On the other hand, the lower capacity firm, Firm 1, earns more than what it would have earned if *retreating* was chosen, since $k_1 < k_2$ requires $\frac{(1-k_1)k_1V}{k_2} > V(1-k_2)$. As expected, after excluding sunk costs, the higher capacity firm makes a higher profit than the lower capacity firm.

Note that the mixed strategy pricing equilibrium bears some resemblance to Chen and Iyer (2002) who find in a model of customized pricing the ratio of profits is equal to the ratio of consumer addressability. In our model, the profit ratio is equal to the ratio of capacities. However, the model in Chen and Iyer (2002) is conceptually very different from ours. In particular, the overlap between the customers of the two firms is always non-zero in Chen and Iyer (2002), whereas in our model the overlap is non-zero only if the sum of the capacities is larger than the market size, i.e., $k_1 + k_2 > 1$. Furthermore, even though the ratio of profits is the same in both papers, the actual profit functions are very different. For example, as mentioned above, and in contrast to Chen and Iyer (2002), the profit of the firm with larger capacity does not depend on its own capacity in our model.

Now suppose $k_1 + k_2 \leq 1$. This implies that each firm that enters the market can sell to its capacity without competing with the other firm for consumers in Segment 3. As such, each firm that successfully enters the market can charge $p_i = V$ and sell k_i units for profit $(V-c)k_i$. Increasing the price will result in zero sales and profit, decreasing the price will still sell k_i units but at lower revenue.

Next consider the capacity subgame equilibrium. The capacity decision is made in anticipation of the possible combinations of values for v_1 and v_2. If both firms find success (i.e., $v_1 = v_2 = V$),

then the profit depends on how k_i and k_j relate to each other and relate to α. The expected profit for Firm i depends on its capacity relative to the capacity of competing Firm j and can be written as follows:

$$E(\pi_i) = \begin{cases} \gamma V k_i - c k_i & \text{if } k_i \leq 1 - k_j \leq 1 - \alpha \\ \gamma(1-\gamma) V k_i + \gamma^2 (\frac{(1-k_i)k_i V}{k_j}) - c k_i & \text{if } 1 - k_j < k_i < k_j \leq 1 - \alpha \\ \gamma(1-\gamma) V k_i + \gamma^2 V (1 - k_j) - c k_i & \text{if } 1 - k_i < k_j \leq k_i \leq 1 - \alpha \\ \gamma(1-\gamma) V (1-\alpha) + \gamma^2 V (1 - k_j) - c k_i & \text{if } \alpha \leq k_j \leq 1 - \alpha < k_i \\ \gamma(1-\gamma) V (1-\alpha) + \gamma^2 V \alpha - c k_i & \text{if } k_i > 1 - \alpha \text{ and } k_j > 1 - \alpha \\ \gamma V (1-\alpha) - c k_i & \text{if } k_i > 1 - \alpha \text{ and } k_j < \alpha \end{cases}$$

where index j indicates the other firm. The equilibrium capacity choices are summarized in the following Proposition.

Proposition 1 *Suppose both firms enter the market initially. The equilibrium capacity choices depend on γ as follows:*

- *If there is a low probability of a successful venture (i.e., $\gamma < c/V$), then both firms choose $k_i = 0$ and earn zero profit.*

- *If there is a moderate probability of a highly successful venture (i.e., $\gamma > c/V$, $\gamma(1-\gamma)V - c > 0$, and $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$), then both firms choose $k_i = 1 - \alpha$ and earn expected profit equal to $(1-\alpha)(V(\gamma(1-\gamma)) - c) + \gamma^2 V \alpha$.*

- *If there is a moderate probability of a moderately successful venture (i.e., $\gamma > c/V$, $\gamma(1-\gamma)V - c > 0$ and $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$), then one firm sets $k = 1-\alpha$ and the other firm sets $k = \frac{\gamma V(1-\alpha(1-\gamma))-c(1-\alpha)}{2\gamma^2 V}$. The higher capacity firm earns expected profit equal to $\frac{c(1-\alpha)+\gamma V(1-\alpha(1-\gamma))}{2}$ and the lower capacity firm earns expected profit equal to $\frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)}$.*

- *If there is a high probability of a successful venture (i.e., $\gamma > c/V$ and $\gamma(1-\gamma)V - c < 0$), then the unique symmetric equilibrium is $k_1 = k_2 = 1/2$. Firms earn expected profit equal to $(\gamma V - c)/2$.*

Proposition 1 highlights a non-monotonic effect of γ on the equilibrium capacity choice. Intuitively, if there is a low probability of success then neither firm wishes to invest in computational capacity because there is a high probability of it going unused. Interestingly, when there is a high probability of a successful venture, firms dampen competition by choosing a capacity that just covers the market. To understand this, consider the extreme case in which $\gamma = 1$. If firms choose computational capacity such that $k_1 + k_2 = 1$, both firms can charge their monopoly price for all of their consumers. The moment capacities are such that firms compete even for a single consumer, the firms are unable to avoid intense price competition for that consumer, thereby affecting revenues from all of their customers. Though an additional unit of capacity can result in an additional sale, the subsequent effect on price competition is severe enough such that firms refrain from competing directly.

Another interesting facet of Proposition 1 is that a moderate probability of a successful venture leads to excessive capacity choices. Therefore, a reduction in the probability of success can actually cause an increase in capacity. To understand this result, consider two competing effects of decreasing γ. On the one hand, lower γ implies greater downside risk that the chosen capacity will go completely unused due to $v_i = 0$ and the resulting failure in the market. This effect would suggest that capacity should decrease as γ decreases. On the other hand, a firm's chance at having monopoly power over all of Segment 3 is maximized at moderate levels of γ. The latter *monopoly harvesting* effect dominates the former *downside risk* effect at moderate levels of γ.

We also see in Proposition 1 that symmetric firms may adopt asymmetric strategies. Since the probability of a successful venture is reasonably high, the potential to serve as a monopolist with probability $\gamma(1-\gamma)$ is lucrative enough such that one firm wishes to expand its capacity beyond the non-competing level of $k_i = 1/2$. However, since the value of being a monopolist is contained (i.e., V is not too high), the second firm loses more in the competitive scenario than it gains in the monopoly scenario by matching the high capacity. As such, one firm will choose capacity to be able to serve all of Segment 3 and the other firm will choose its best response recognizing the drawback of maximum capacity in the case of competition.

The results of Proposition 1 are depicted in Figure 5.[9] Note that Region 2 in which both firms

[9]In Figures 5–9, we use parameters $\alpha = \frac{1}{4}$, $c = \frac{1}{2}$, $V = \frac{7}{2}$ and $F = 0$, unless that parameter is being used as a variable in the figure.

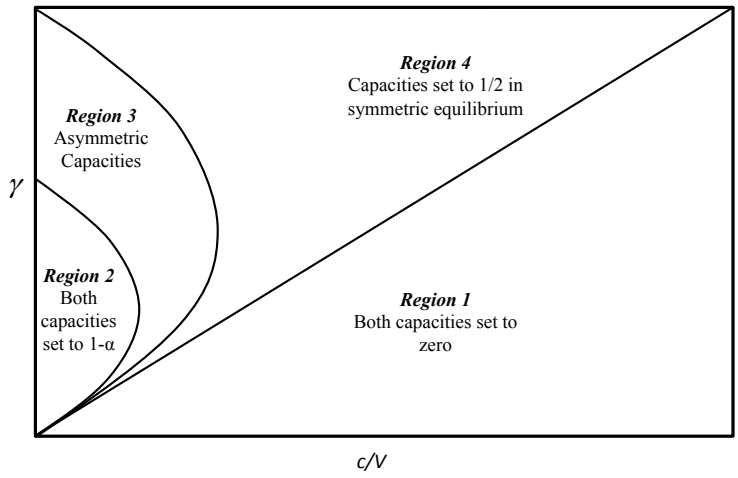

Figure 5: Equilibrium capacities as a function of γ and c/V

set capacity $k = 1-\alpha$ disappears when $\alpha = 0$. In other words, as the firms become undifferentiated in the minds of consumers, they will either differentiate in their capacity choice (Region 3) or dampen competition via restricted capacity choices (Region 4). As α increases, Region 2 expands at the expense of Region 3 in which firms choose asymmetric capacities. In fact, the border between Region 3 and Region 4 in which both firms choose $k = 1/2$ becomes the border between Region 2 and Region 4 as α approaches $1/2$, thereby eliminating Region 3. In other words, as the firms become more differentiated in the minds of consumers, they become less likely to adopt asymmetric strategies.

We now turn our attention to the entry decision. If only one firm enters the market, the firm will be a monopolist optimally choosing $k = 1 - \alpha$ and earning expected profit $(\gamma V - c)(1 - \alpha)$ if $\gamma > c/V$ and optimally choosing $k = 0$ to earn zero profit otherwise.

Lemma 1 *If $F > (\gamma V - c)(1 - \alpha)$, then there is no entry in the conventional model. Otherwise, the entry decisions depend on F, γ, and V as follows:*

- *If there is a moderate probability of a highly successful venture (i.e., $\gamma > c/V$ and $\gamma(1 - \gamma)V - c > 0$ and $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$), both firms enter if $F < (1 - \alpha)(V\gamma(1 - \gamma) - c) + \gamma^2 V\alpha$. Otherwise only one firm enters.*

- *If there is a moderate probability of a moderately successful venture (i.e., $\gamma > c/V$ and $\gamma(1 -$*

$\gamma)V - c > 0$ *and* $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$*), both firms enter if* $F < \frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)}$*. Otherwise only one firm enters.*

- *If there is a high probability of a successful venture (i.e.,* $\gamma > c/V$ *and* $\gamma(1-\gamma)V - c < 0$*), then both firms enter if* $F < (\gamma V - c)/2$*. Otherwise only one firm enters.*

Lemma 1 shows how the probability of success can affect the firms' decision to enter. Next, we consider the effect of autoscaling on firms' entry decisions and compare it to our findings from this Lemma.

3.3.2 The Effects of Autoscaling

We now turn our attention to the equilibrium when autoscaling is available. We allow for firms to decide whether to use autoscaling or prepay for computational capacity after the entry decision. Major cloud providers offer autoscaling with no additional fees.[10] We first solve for the equilibrium pricing supposing both firms enter and choose autoscaling. We will then look at the equilibrium pricing supposing only one firm chooses autoscaling to identify the equilibrium decisions of whether to adopt autoscaling. Upon identifying the equilibrium adoption of autoscaling conditional on entry, we look at how firms' prices, profits, and entry decisions are affected by autoscaling.

First, consider the case where both firms enter and choose autoscaling. With probability γ^2, we have $v_1 = v_2 = V$, and it is straightforward to show there is no pure strategy pricing equilibrium; instead the pricing subgame leads to a mixed strategy equilibrium where the prices of both firms range between z' and V. Similar to our analysis of mixed strategy equilibrium without autoscaling, z' is the price for which each firm is indifferent between *attacking*, resulting in a profit of $(z'-c)(1-\alpha)$, and *retreating*, resulting in a profit of $\alpha(V-c)$. This results in $z' = \frac{\alpha(V-c)}{1-\alpha} + c$.

Supposing that $G_i(.)$ is the cumulative distribution function for the price of Firm i, the profit of Firm i when setting price x, is

$$\pi_i(x) = \alpha G_j(x)(x-c) + (1-G_j(x))(1-\alpha)(x-c)$$

Setting the derivative of this function equal to zero for $x \in (z', V)$ and using the boundary conditions

[10] https://aws.amazon.com/autoscaling/pricing/, accessed September 2016.

$G(z') = 0$ or $G(V) = 1$, we find

$$G_j(x) = \begin{cases} 0 & \text{if } x < z' \\ \frac{(1-\alpha)(x-c)-\alpha(V-c)}{(1-2\alpha)(x-c)} & \text{if } z' \leq x \leq V \\ 1 & \text{if } x > V \end{cases}$$

which results in the expected profit $\pi_{AA} = \alpha(V-c)$ for each firm, when they both use autoscaling.

With probability $\gamma(1-\gamma)$, $v_1 = V$ and $v_2 = 0$, giving Firm 1 monopoly power over all of Segment 3 and profit of $(V-c)(1-\alpha)$. Thus, expected profit with autoscaling when both firms enter is $(V-c)(\gamma^2\alpha + \gamma(1-\gamma)(1-\alpha))$.

Finally, consider the case where both firms enter but only one firm adopts autoscaling. Without loss of generality, we assume that Firm 2 adopts autoscaling and Firm 1 chooses capacity k_1. Using the same techniques as in Section 3.3.1, we prove in the Technical Appendix that there is no pure strategy equilibrium for prices. The mixed strategy, represented by cumulative distribution function H_i, is given by

$$H_1(x) = \begin{cases} 0 & \text{if } x < z'' \\ \frac{-ck_1+(-1+k_1)V+\alpha(c-x)+x}{(k_1-\alpha)(x-c)} & \text{if } z'' \leq x < V \\ 1 & \text{if } x \geq V \end{cases}$$

$$H_2(x) = \begin{cases} 0 & \text{if } x < z'' \\ \frac{k_1(-ck_1+(-1+k_1)V+\alpha(c-x)+x)}{(-1+\alpha)(\alpha-k_1)x} & \text{if } z'' \leq x < V \\ 1 & \text{if } x \geq V \end{cases}$$

where $z'' = \frac{\alpha c - ck_1 + (k_1-1)V}{\alpha - 1}$ is the price at which Firm 1 is indifferent between attacking at price z'' and retreating at price V. Note that $H_2(x)$ is discontinuous at $x = V$. This implies that Firm 2 uses price V with probability $1 - \frac{k_1(V-c)}{(1-\alpha)V}$. In other words, $h_2(V) = (1 - \frac{k_1(V-c)}{(1-\alpha)V})\delta(0)$, where $h_2(.)$ is the probability density function for price of Firm 2 and $\delta(.)$ is Dirac delta function. Prior to the pricing game, the optimal capacity k_1 that maximizes expected profit for Firm 1 is given by:

$$k_1^* = \frac{V\gamma(1-\alpha+\alpha\gamma) - c\left(1+\alpha\left(-1+\gamma^2\right)\right)}{2(V-c)\gamma^2}$$

We may now examine the equilibrium adoption of autoscaling. The payoffs from each possible firm choice of autoscaling or capacity k are summarized in Table 2.

	Firm 2 uses Autoscaling	**Firm 2 uses capacity k**
Firm 1 uses Autoscaling	$\pi_1 = (V-c)(\gamma^2\alpha + \gamma(1-\gamma)(1-\alpha))$ $\pi_2 = (V-c)(\gamma^2\alpha + \gamma(1-\gamma)(1-\alpha))$	$\pi_1 = \gamma^2(1-k)(v-c) + \gamma(1-\gamma)(1-\alpha)(v-c)$ $\pi_2 = \gamma^2\frac{k(c(k-\alpha)+V(1-k))}{1-\alpha} + \gamma(1-\gamma)kV - ck$
Firm 1 uses capacity k	$\pi_1 = \gamma^2\frac{k(c(k-\alpha)+V(1-k))}{1-\alpha} + \gamma(1-\gamma)kV - ck$ $\pi_2 = \gamma^2(1-k)(v-c) + \gamma(1-\gamma)(1-\alpha)(v-c)$	Profits are the same as in Section 3.3.1

Table 2: Payoffs from autoscaling adoption strategies assuming both firms enter the market

Comparing payoffs across cases in Table 2, the equilibrium adoption of autoscaling is summarized in the following lemma.

Lemma 2 *A firm's decision on using autoscaling depends on the number of market entrants, α, γ, c, and V as follows:*

- *If both firms enter the market, both will choose autoscaling for low enough probability of success, $\gamma < (1-\alpha)/(2-3\alpha)$, or high enough computational costs, such that $c/V > L$;*
- *If both firms enter the market, only one firm will choose autoscaling for high enough probability of success and low enough computational costs, such that $\gamma > (1-\alpha)/(2-3\alpha)$ and $c/V < L$;*
- *If one firm enters the market, it will choose autoscaling;*

where L is defined as:

$$L \equiv \frac{\gamma^2(-3\alpha\gamma + \alpha + 2\gamma - 1)^2}{2(1-\gamma)\sqrt{(1-\alpha)^3\gamma^3(2\alpha\gamma - \alpha - \gamma + 1)} + \gamma\left((2-3\alpha)^2\gamma^3 + ((9-5\alpha)\alpha - 4)\gamma^2 - (\alpha-1)\alpha\gamma + (\alpha-1)^2\right)}$$

Lemma 2 establishes the equilibrium adoption of autoscaling, which will be used in subsequent analysis of the effect of autoscaling's availability on prices, entry, profits, and consumer surplus. Autoscaling allows the firms to avoid over- or under-spending on computational capacity. However, there is also a strategic effect of autoscaling that can lead to dampened or intensified competition. Interestingly, the Lemma shows that symmetric firms may make asymmetric decisions regarding the use of autoscaling. When competition is sufficiently intensified by autoscaling, one firm chooses a fixed capacity in order to alleviate competition to some extent.

Effect of Autoscaling on Equilibrium Prices

Given the firms' equilibrium strategies, we can examine how autoscaling will affect equilibrium prices in the event that entry costs are low enough such that multiple firms enter.

Proposition 2 *Suppose entry costs are such that both firms enter the market with or without autoscaling. The effect of autoscaling on average prices depends on γ as follows:*

- *If the probability of a successful venture is not too high (i.e., $\gamma(1-\gamma)V - c > 0$ and $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$), then autoscaling increases the average price set by each firm.*
- *If the probability of a successful venture is high enough (i.e., $\gamma > c/V$ and $\gamma(1-\gamma)V - c < 0$), then autoscaling decreases the average price set by each firm.*

Proposition 2 shows the effect of autoscaling on the average prices charged by each firm in Regions 2 and 4 of Figure 5. In Region 2, both firms would choose enough capacity to serve their own segment (Segment i for Firm i) and all of Segment 3 without autoscaling. The cause of the change in average prices with the addition of autoscaling is that autoscaling turns the cost of each server from a sunk cost to a cost that depends on the number of consumers served by the firm. Without autoscaling, firms do not consider the cost of servers in their pricing decision, as this cost is sunk. Thus, they receive no negative utility from serving a larger portion of the market and are more flexible to do so by decreasing price. However, with autoscaling, each additional customer adds an additional cost, resulting in firms having less incentive to decrease their price to get more customers compared to when costs were sunk. Therefore, autoscaling decreases the firms' incentive to *attack* aggressively with price and increases the average price in Region 2. Thus demand uncertainty creates an important distinction from previous literature on capacity choice (e.g., Kreps and Scheinkman 1983), as it results in a capacity game that *decreases* average prices relative to the pricing game that arises with autoscaling.

On the other hand, the price-competition dampening effect of autoscaling is mitigated in Region 4 where firms would choose not to *attack* without autoscaling since they restrict their capacity to dampen competition. The equilibrium choice of capacity results in both firms charging the maximum price of V. Autoscaling removes this separation of targeted consumers and increases competition between the two firms, resulting in decreased average prices in Region 4.

Proposition 2 thus shows that the probability of a successful venture is critical in determining whether autoscaling increases or decreases prices; a result that is new to the literature.

Note that Region 3 in Figure 5 represents a case where the two firms have different average prices without autoscaling. Since with both firms using autoscaling they both set the same price, evaluating the effect of autoscaling on average charged prices is not as straight forward as in Regions 2 and 4. Later in this section, we use consumer surplus as a proxy to average prices to study the effects of autoscaling in this region. But first, we examine the firms' entry decisions.

Effect of Autoscaling on Entry Decisions

Given the firms' equilibrium strategies and their expected profits, we can derive their entry decisions. We summarize the entry decisions with autoscaling in the following lemma.

Lemma 3 *If* $F > \gamma(V-c)(1-\alpha)$*, then there is no entry in the autoscaling model. Both firms enter the market if* $F < Max[(V-c)(\gamma^2\alpha+\gamma(1-\gamma)(1-\alpha)), \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma v(\alpha(-\gamma)+\alpha-1)\right)^2}{4(\alpha-1)\gamma^2(c-V)}]$*. Otherwise, only one firm enters.*

By comparing the results presented in Lemma 1 to Lemma 3, we can determine the effect of autoscaling on firm entry. We first consider the effect of autoscaling on participation in the market by any of the firms, i.e., how autoscaling affects whether at least one of the two firms enters.

Proposition 3 *Autoscaling increases the range of entry costs,* F*, such that at least one firm enters the market.*

Proposition 3 confirms common intuition that autoscaling can make entry more attractive for at least one firm. The reason is that it allows firms with uncertain likelihood of success to incur the cost of computational needs after demand is realized. This highlights the *downside risk reducing* effect of autoscaling. Without autoscaling, firms have to invest in the cost of entry F and the cost of computational capacity ck_i prior to realizing whether the venture will be successful. Autoscaling increases the range of entry costs for which at least one firm enters by $c(1-\gamma)(1-\alpha)$. Thus, the higher the cost of capacity and the lower the probability of success, the more effective autoscaling will be in guaranteeing that the market will be served by at least one firm. We now examine how autoscaling affects whether multiple firms enter the market.

Proposition 4 *There exists a cut-off $\hat{\gamma}$ for which autoscaling decreases the range of entry costs, F, such that both firms enter if and only if $\gamma > \hat{\gamma}$.*

Proposition 4 finds the counter-intuitive result that autoscaling can decrease market entry. Though autoscaling has a *downside risk reducing* effect, it also has a *competition intensifying* effect. In other words, autoscaling makes it less costly for a firm to find out if it has a successful venture on its hands, but also less costly for a firm to fight aggressively for consumers in Segment 3. To further explain these effects and when each is dominant, we consider the three potential outcomes if both firms enter.

When both firms enter, there is a $\gamma(1-\gamma)$ probability that a firm finds itself a monopolist, a γ^2 probability that a firm finds itself competing head-to-head, and a $1-\gamma$ probability that a firm finds $v_i = 0$. In the former case, autoscaling weakly benefits firms because they are assured of having the computational capacity to satisfy the demand of all consumers in Segment 3. Without autoscaling, firms acknowledging the downside risk choose $k_i \leq 1-\alpha$ and thus cannot satisfy all demand when given monopoly power over all consumers in Segment 3. This is the problem that startup Befunky experienced without autoscaling in the earlier example and represents a positive *demand satisfaction* effect of autoscaling. In the latter case, autoscaling weakly benefits firms because it prevents them from over-purchasing capacity. Without autoscaling $k_i \geq 0$ and thus firms have excess computational capacity when $v_i = 0$. This is the positive *downside risk reducing* effect of autoscaling. However, autoscaling weakly disadvantages firms if they compete head-to-head. The same fact that benefits firms when they are monopolists is damaging when they compete: with autoscaling firms are assured of having computational capacity to satisfy demand of all consumers in Segment 3. As such, autoscaling intensifies competition. Without autoscaling, the fact that $k_i \leq 1-\alpha$ allows firms to include a more profitable *retreating* price in the equilibrium mixed strategy. In this case, the demand satisfaction effect actually leads to the negative *competition intensifying* effect of autoscaling.

The *downside risk reducing* effect is most dominant when γ is low. The *demand satisfaction* effect is most dominant when $\gamma(1-\gamma)$ is high (i.e., moderate γ) and the *competition intensifying* effect is most dominant when γ is high. A high γ increases the probability of competition and also makes it such that firms without autoscaling choose capacities such that this competition is

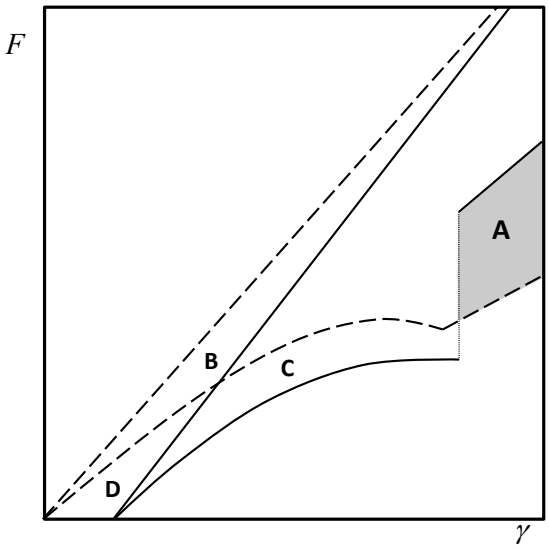

Figure 6: The effect of autoscaling on entry: Region A denotes autoscaling decreases entry from 2 to 1 firms. Region B denotes autoscaling increases entry from 0 to 1 firm. Region C denotes autoscaling increases entry from 1 to 2 firms. Region D denotes autoscaling increases entry from 0 to 2 firms.

avoided. Therefore, autoscaling decreases market entry when the probability of a successful venture is sufficiently high. Propositions 3 and 4 suggest that to find the effect of autoscaling on the number of new entrants, we must consider the probability of success as well as the cost of entry, which goes against the intuition that autoscaling always facilitates market entry.

The results of Propositions 3 and 4 are graphically depicted in Figure 6. As shown in this figure, there are four regions of interest. In region A, autoscaling decreases entry due to the *competition intensifying* effect. Autoscaling allows one firm to be a monopolist because the other firm cannot profitably enter given the anticipated level of competitive intensity. In regions B and D, the market will not be served by either firm unless there is autoscaling. In region C, a firm would have monopoly power because the downside risk of capacity pre-purchase makes it unprofitable for a second entrant, but autoscaling alleviates these effects and results in competing firms entering the market.

Corollary 1 *For low costs of entry, autoscaling can create a prisoner's dilemma. Both firms choose autoscaling even though they earn greater expected profit in the absence of autoscaling.*

As noted previously, the *competition intensifying* effect can outweigh the *demand satisfaction*

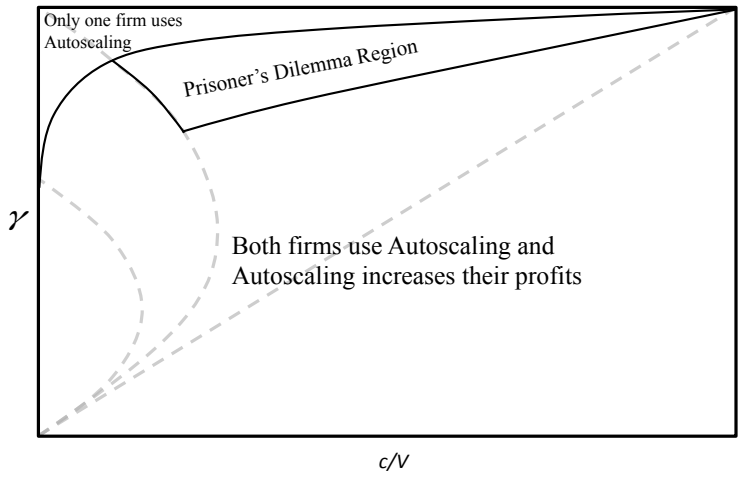

Figure 7: When existence of autoscaling can lead to a prisoner's dilemma effect for the firms

effect and the *downside risk reducing* effect for sufficiently high γ. If F is sufficiently low, both firms will choose to enter with or without autoscaling. Furthermore, as shown in Corollary 1, they both choose autoscaling in equilibrium. Interestingly, this leads to a prisoner's dilemma situation where the firms' adoption of autoscaling results in diminished expected profitability of both firms. This result is depicted in Figure 7. The dashed lines in Figure 7 correspond to regions when autoscaling is not available (from Figure 5), and show how autoscaling affects firms' equilibrium profits in different regions. When the probability of success, γ, is very high, only one firm uses autoscaling while the competing firm can strategically limit its computational capacity to soften competition. Also, when γ is sufficiently low, both firms use autoscaling, but due to the downside risk reducing effect of autoscaling, both firms get higher profits with autoscaling. However, a moderately high γ creates a prisoner's dilemma situation where the competition intensifying effect of autoscaling dominates the downside risk reducing effect, but the firms still use autoscaling. Therefore, both firms would be better off if autoscaling was not available in this region. We now turn our attention to the impact of autoscaling on consumers.

Effect of Autoscaling on Consumer Surplus

So far, we studied the effects of autoscaling on firms' strategies and their profit. Autoscaling can increase price competition between firms. It can also increase market entry. Both of these

effects, intuitively, should lead to higher surplus for consumers. However, autoscaling also changes the capacity cost from sunk cost at the time of pricing to variable cost. Therefore, as shown in Proposition 2, autoscaling can lead to higher average prices, and thus lower surplus, for consumers. Furthermore, as shown in Proposition 4, autoscaling can also increase the likelihood of a monopoly market. In this section, we study the effect of these opposing forces on consumer surplus.

Expected consumer surplus can be derived from calculating the difference between expected social welfare and combined expected firm profit. Social welfare is equal to the combined value consumers get (i.e., V times the number of purchases) minus the cost to deliver that value (i.e., c times the computational capacity). Therefore, the expected consumer surplus can be written as

$$E[CS] = V \times E[\#\text{purchases}] - c \times E[\text{computational capacity}] - E[\pi_1] - E[\pi_2] \tag{3}$$

where #purchases and computational capacity indicate the total number of consumers who purchase the product and the total computational capacity reserved by the firms, respectively. If there is only one firm in the market, in both cases with and without autoscaling, that firm sets the price to V and results in zero consumer surplus. If both firms are in the market, consumer surplus with autoscaling available depends on whether both firms choose to adopt autoscaling or if only firm adopts this feature. We present the values of consumer surplus derived from Equation (3) in the Technical Appendix. Comparing across conditions, we have the following result.

Proposition 5 *For sufficiently small F, sufficiently small c and a moderate value of γ, autoscaling decreases expected consumer surplus. If F is sufficiently large, autoscaling has no effect on the expected consumer surplus. Otherwise, autoscaling increases the expected consumer surplus.*

Proposition 5 shows the effect of autoscaling on consumer surplus. The result is also depicted in Figures 8 and 9. Figure 8 is analogous to Figure 5 with the contours from Figure 5 marked with dashed gray lines. As we can see, for Region 2 of Figure 5, where both firms set their capacities to $1 - \alpha$ in the absence of autoscaling, and part of Region 3, where firms set asymmetric capacities, autoscaling decreases consumer surplus. Intuitively, since autoscaling increases average prices in this region, it decreases consumer surplus.

Figure 9 is analogous of Figure 6 with the contours from Figure 6 marked as dashed gray lines. The figure shows that autoscaling can only affect consumer surplus if the cost of entry is

sufficiently low. In particular, unless both firms enter the market in at least one of the two cases (i.e., autoscaling or no autoscaling available), consumer surplus will be zero; therefore, if F is not low enough, autoscaling can have no effect on consumer surplus. Figure 9 also shows that when autoscaling increases market entry to two firms, it also increases expected consumer surplus.

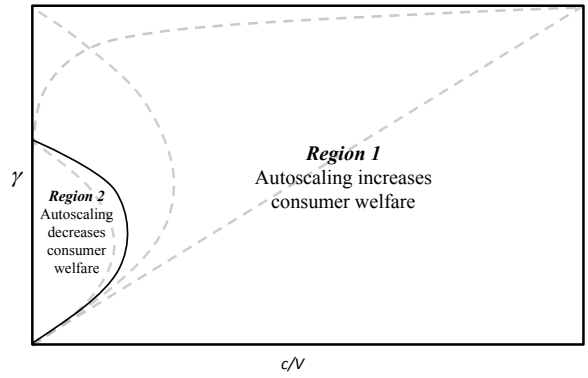

Figure 8: Effect of autoscaling on consumer surplus as a function of V, c and γ

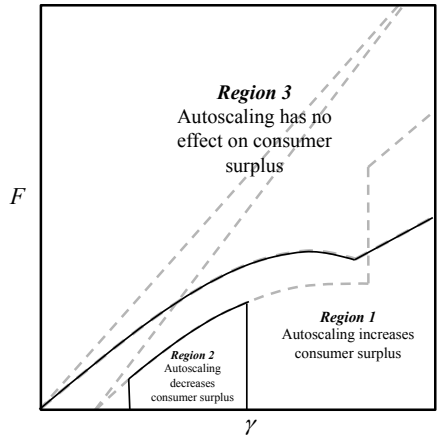

Figure 9: Effect of autoscaling on consumer surplus as a function of F and γ

3.4 Discussion

The emergence of cloud computing and its feature of autoscaling has the potential to influence the entry of startups into new markets. Before autoscaling was available, startups needed to invest heavily in computational capacity before entering the market so that they could serve highly unpredictable demand levels. The uncertainty in demand meant that the capacity chosen by firms could be more or less than the actual needs, resulting in extra costs or unfulfilled demand. However, autoscaling allows the new entrants a flexible capacity such that the cloud provider will designate the specific computational resources as the actual demand and required resources are realized. This feature can help new entrants to the market pay only for the capacity that is required to meet their demand, reducing the disadvantages of uncertainty in demand. In this paper, we find the effects of autoscaling on entry decisions of startups, their prices, their capacity decisions, and their profits.

Our model shows how the *demand satisfaction* effect of autoscaling, in which startups can be assured of having capacity to satisfy demand, turns out to be a double-edged sword in that it frees competing startups to aggressively pursue customers. Our research identifies this *competition*

intensifying effect of autoscaling and establishes the conditions under which this effect will outweigh the positive effects of autoscaling. This effect has several important implications for startups.

The results can help guide pricing decisions by startups who enter the market with autoscaling available. We find that autoscaling will increase the average prices startups charge if the probability of a successful venture is not too high. In this case, autoscaling decreases a startup's costs, but these savings are not passed on to consumers because autoscaling converts a sunk fixed cost into a variable cost that a startup incurs if and only if they attract more customers. As a consequence, price competition is dampened by autoscaling in the case that startups would otherwise choose excessive capacities. However, if the probability of a successful venture is sufficiently high, startups would optimally limit their capacities to dampen competition in the absence of autoscaling. In this case, autoscaling gives the startups freedom to aggressively pursue customers with price and thus can decrease average prices.

The trade-off between the intensified competition and the reduced downside risk of failure affects startups' decisions on whether to enter the market. In particular, when the probability of success is sufficiently high, the increased competition effect of autoscaling dominates the decreased downside risk of failure. As a result, while autoscaling facilitates the entry for the first startup, it deters the second startup from entering the market. Our findings expand the market entry literature by looking at the topic from the new angle of capacity choice before versus after demand is realized.

Autoscaling could also have positive or negative effects on consumer surplus. On the one hand, it can increase price competition between the firms and facilitate their entry, both of which lead to higher consumer surplus. On the other hand, autoscaling can lower consumer surplus by dampening price competition or by discouraging the second startup from entering the market if one startup has already entered. Our results show that when cost of entry is low and probability of success is moderate, autoscaling decreases expected consumer surplus.

Overall, our results highlight how startups should use information on cost of entry, probability of success, cost of capacity and level of differentiation to evaluate their entry decisions, capacity commitments and pricing strategies in the presence of autoscaling. Our research is one of the first to consider the marketing aspects of cloud computing and autoscaling. With the rapid adoption of cloud computing by firms across different industries, marketing and economics research on the cloud can be a rich and important topic of study for future research.

4 Essay 2: Spot Pricing in Cloud Computing

4.1 Introduction

Spot pricing allows firms access to cost effective cloud computing resources by putting the cloud provider's excess resources on sale. As demand for cloud computing resources fluctuates, cloud providers often have temporary unused resources, which can be offered on the spot market at a discount. Spot prices can be 90% lower than regular on-demand prices. Thus, using spot resources, companies can significantly reduce their computation costs. For instance, IronSource has been abale to save up to 80% on computation costs using spot pricing for scaling their big data solutions in periods of high demand. Similarly, the ride-sharing company, Lyft, has reported a 90% monthly saving after using Amazon's AWS spot pricing[11].

In contrast to On-Demand pricing, which offers users access to computing resources over a time period for a fixed price, with spot pricing there is no guarantee over how long the excess resources remain available. Spot resources may become unavailable after purchase if they are requested by On-Demand users, who are prioritized over spot users when resources are scarce. Therefore, while spot prices are generally much lower than On-Demand prices, there is more uncertainty on whether a spot instance will remain uninterrupted throughout its intended duration. Moreover, while on-demand prices are kept unchanged, spot prices are subject to fluctuation as demand for a particular cloud resource changes.

In order to alleviate the effects of interruptions for spot users, cloud computing providers offer the option to diversify cloud resources when using spot pricing. This option allows spot users to use resources from a variety of different server types, instead of only using from one pool of servers. For instance, Amazon Web Services allows spot users the option to choose between two strategies: Minimize and Diversify[12]. The Minimize strategy allows users to host all their resources from the same pool of servers, choosing the lowest priced pool. The Diversify strategy allows users to spread their resources across multiple pools of servers. One benefit of a diversified portfolio of cloud resources is that if an interruption terminates one pool of servers, the other instances can continue running without interruption.

[11] https://aws.amazon.com/ec2/spot/testimonials/

[12] https://aws.amazon.com/about-aws/whats-new/2015/09/amazon-ec2-introduces-diversified-and-lowest-priced-spot-instance-fleets/

Cloud computing resources, also referred to as instances, are categorized into different pools, based on their type and geographical region. There can be multiple sources of interruption in a spot instance from a particular pool. Increasing demand for a certain pool of instances can result in high prices that are above the user's value for the instance. Instances from one pool can be pulled from the spot market if they are requested by on-demand users. Finally, macro level interruptions in the cloud provider's servers, such as unanticipated natural disruptions or mistakes from employees, can make the pool unavailable. Diversifying cloud resources across multiple pools can thus reduce the number of a user's terminated instances, in case of an interruption in one instance pool. Therefore, it is commonly believed that implementing a diversified cloud strategy is justified primarily when interruptions are highly likely. However, this notion does not take into consideration the pricing decisions of the cloud providers. Cloud providers set prices for each pool based on the supply and demand of instances and change prices as these factors change. By incorporating the strategic pricing decision of cloud providers in users' decisions for diversifying, we show that, despite common intuitions, diversifying can actually be more beneficial in cases when interruptions are unlikely to occur.

Specifically, we find the effect of interruption likelihood on diversification depends on the value difference between users in the spot market. Consider two users, one with high valuation and the other with low valuation for cloud instances. Our findings show when the value difference between the two users is low, diversification occurs for high interruption likelihood. However, when the value difference is moderately high, the reverse occurs and low interruption likelihood results in diversification. This finding contradicts the notion that diversification should always be chosen as insurance against interruptions.

We model two pools of instances and two users, each of whom demand two instances. When one user diversifies, there is a shared pool hosting instances from both users. This shared pool becomes the only available pool, if an interruption occurs in the other pool. In this case, if the value difference between the users is high, the high value user is a much more attractive customer and the provider finds it optimal to set a high price for the remaining pool only selling to the high value user, instead of selling the shared pool to both users at a lower price. Without an interruption, the provider could charge a high price for one of the high value user's instances in one pool and charge a low price in the shared pool to sell to both users. Thus interruptions can

increase the price of the remaining pool for a diversifying high value user, if the value difference between users is high. Therefore, for high value difference, diversification is only chosen for low interruption likelihoods.

We also show that when the value difference between the two users is low and the high value user diversifies, an interruption in the pool not shared with the low value user can actually benefit the high value user. This means that while an increase in interruption likelihood does not benefit a user, the actual occurrence of an interruption may be beneficial for high value users. Without an interruption, the high value user buys two instances, one at a high price in a pool dedicated to the user and the other at a low price in a pool shared with the low value user. The marginal utility of the additional second instance is not as much as the utility from the first instance for the user. As the high value user's dedicated pool is interrupted, the user only buys one instance but at the lower price of the shared pool. Thus, the average price per instance goes down for the high value user as interruption occurs, making the user better off.

This paper provides insights for both users and providers of cloud computing. Based on our findings, low value users are better off not diversifying. For high value users, the choice of diversification depends on the value difference among users and the likelihood of interruption. Using a model of the provider's pricing strategy, we show that interruptions are not always damaging to users, as they may reduce average prices. We also find how diversification choices affect cloud providers' profit. Theoretically, this study contributes to the literature on cloud computing, as well as the price discrimination literature. The rest of this document is organized as follows. First, we introduce the model. Next, we analyze the model and present the results. Finally, we provide a discussion of the findings.

4.2 Model

We consider one monopolist cloud provider offering instances from two different pools on the spot market. The likelihood that each pool is interrupted and becomes unavailable is γ, where $0 < \gamma < 1$. We assume there are no limits to the number of instances that can be provided within each pool, as long as no interruptions occur. Once a pool is interrupted, the entire pool becomes unavailable. In this model, γ is determined by forces outside of the spot market, thus representing exogenous macro level sources of interruption (e.g. allocation of the pool to on-demand users or natural causes

of disruptions).

There are two users, a high value firm and a low value firm. To capture the choice of diversification, we assume each user has value for two instances. The marginal value of an additional instance is lower than the value of the first instance by a factor of $\alpha < 1$. Cloud computing resources are used for applications with varying degrees of tolerance for delay (Yi et al. 2016). This means while some computation tasks can be interrupted without a huge loss, other asks are more time sensitive and are thus costlier to interrupt. Testimonials from spot market users demonstrate that applications range from interruption tolerant tasks (e.g. Novartis using spot instances to conduct data analysis and research) to highly time sensitive tasks (e.g. ironSource using spot instances to handle surges in consumer demand)[13]. This variance implies that the cost of an interruption in a computation task depends on how critical the task is to the user. Therefore, we assume that not all tasks are valued the same by a firm. Thus, if a firm has access to only one instance, it would run its most critical and high-value application. If an additional instance becomes available, a task with a lower priority and value will be added. Therefore, we assume the high value user's value v_h for the first instance and αv_h for the second instance. Similarly, the low value user's value v_l for the first instance and αv_l for the second instance, where $v_l < v_h$.

Users submit a request on how they wish to allocate their two instances among the two pools. There are two allocation strategies: *Diversify* and *Minimize*. This reflects the choices given to users by spot providers such as Amazon[14]. If they choose to Diversify, they request for each of their instances to be launched in a different pool. Otherwise, they choose to Minimize and launch both instances in a single pool (note that the instances within each pool are priced similarly).

The timing of the game is as follows. With common knowledge of the interruption probability, users first submit their allocation requests (i.e. Diversify or Minimize). At this point, the actual available supply of instances is unknown because the decision is made prior to any potential interruptions being realized. The assumption that allocation requests are made first captures the reality that when users launch spot instances, they are uncertain about the future availability and price of the instances, as spot prices and pool availability are subject to changes and can be highly unpredictable[15].

[13] https://aws.amazon.com/ec2/spot/testimonials/
[14] https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet.html#spot-fleet-allocation-strategy
[15] https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances-history.html

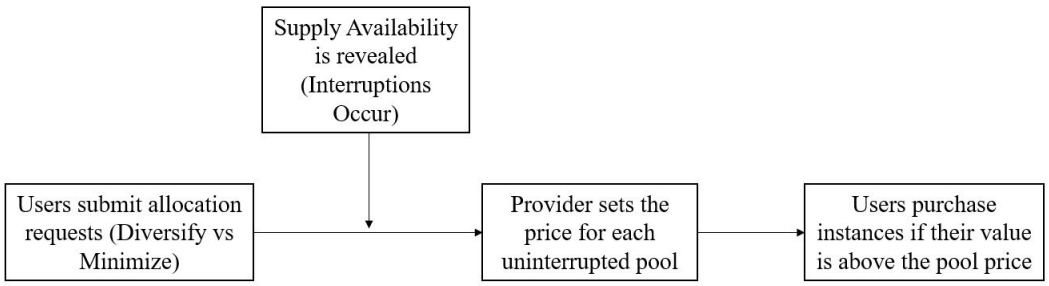

Figure 10: Timing of Decisions in the Model

In the next stage, the availability of supply is revealed, as nature determines whether an interruption occurs in each pool according to the probability γ. Then observing the users' allocation requests and whether or not there interruption in each pool, the provider sets prices for each pool, p_1 and p_2. Note that after the realization of supply, the provider is not constrained to keep the price for the users requesting Minimize allocation at the minimum. Choosing a Minimize allocation does not come with a guarantee that the price of instances will remain unchanged after supply is known, only that all requested instances will be priced the same. In fact, Amazon notifies users that when using Minimize allocation strategies, their cost can become "unexpectedly high if pricing in the pool spikes"[16]. Finally, users purchase each requested instance only if their value for the instance exceeds the price of its pool[17]. The order of decisions is shown in Figure 10. Table 3 summarizes the notations used in the model.

Symbol	**Description**
v_l	Low value user's valuation for its first instance i
v_h	High value user's valuation for its first instance i
α	Relative value of the second instance compared to the first instance
γ	Likelihood of interruption in each pool
A	Available pool
I	Interrupted pool
p_i^{AI}	Price of pool i, when the first pool is interrupted (I) and the second pool is available (A)
π_i^{AI}	Provider's profit in pool i, when the first pool is interrupted and the second pool is available
$uLow_{MM}^{AA}$	Low value user's utility when both firms minimize (MM) and both pools are available (AA)
$uHigh_{DD}^{AA}$	High value user's utility when both firms diversify (DD) and both pools are available (AA)

Table 3: Timing of Decisions in the Model

[16] https://aws.amazon.com/blogs/aws/new-spot-fleet-option-distribute-your-fleet-across-multiple-capacity-pools/

[17] This assumption captures the fact that firms can submit a maximum price for each request, which means when a server's price is set higher than a user's maximum value, the server will not be sold to that user. Once a contract is approved, the continued availability of a server is not guaranteed by the provider. If a server becomes unavailable, the user will of course not have to pay for it any longer.

4.3 Analysis

In the analysis below, we use backward induction to derive optimal allocation choices. We begin by finding the optimal pool prices that maximize provider profits given each particular allocation strategy and interruption occurrence. For each set of allocation decisions, the prices depend on which pools are available. Next, we find the expected payoffs for each user from each set of allocation choices, by taking into account the likelihood that interruptions may occur. Finally, comparing expected user payoffs from each allocation strategy, we find optimal allocation strategies as a function of the model parameters. With each user choosing between diversifying and minimizing, there are four possible sets of allocation strategies analyzed below.

4.3.1 Both Users Minimize

Suppose both pools are available and the two users' instances are hosted in separate pools. We begin deriving payoffs and prices after interruptions are known. The provider's prices are set after interruptions occur, creating the following cases.

(I) Both pools are available

The provider has two pricing options for the high value user's pool: 1) selling two instances, each at the price of αv_h, 2) selling one instance at the price of v_h. Similarly, the provider has two pricing options for the low value user's pool: 1) selling two instances, each at the price of αv_l, 2) selling one instance at the price of v_l.

Thus, the provider's profit is $Max[2\alpha v_h, v_h]$ from the high value user and $Max[2\alpha v_l, v_l]$ from the low value user. We assume from here on that $\alpha > \frac{1}{2}$, so that it is optimal for the provider to sell both instances to a user in the case that there is only one user assigned to a given pool. When each user is hosted on a separate pool, we have $2\alpha v_h > v_h$ and $2\alpha v_l > v_l$. Thus, high value user's pool is priced at αv_h and low value user's pool is priced at αv_l. The provider's profit without interruption is $\pi_{MM}^{AA} = 2\alpha(v_h + v_l)$, where the subscript MM represents both users minimizing and the superscript AA represents both pools being available.

The high value user's utility, given both pools are available (AA), is $uHigh_{MM}^{AA} = (v_h + \alpha v_h) - 2\alpha v_h = v_h(1-\alpha)$. The low value user's utility, given both pools are available (AA), is $uLow_{MM}^{AA} = (v_l + \alpha v_l) - 2\alpha v_l = v_l(1-\alpha)$.

(II) High value user's pool is interrupted and low value user's pool is available

Low value user's pool is priced at αv_l and the provider's profit is $\pi^{IA}_{MM} = 2\alpha v_l$, where the subscript MM represents both users minimizing and the superscript IA represents the high value user's pool being interrupted and the low value user's pool being available. The low value user's utility, given an interruption in the high value user's pool, is $uLow^{IA}_{MM} = (v_l + \alpha v_l) - 2\alpha v_l = v_l(1-\alpha)$.

(III) Low value user's pool is interrupted and high value user's pool is available

High value user's pool is priced at αv_h and the provider's profit is $\pi^{AI}_{MM} = 2\alpha v_h$. High value user's utility is $uHigh^{AI}_{MM} = (v_h + \alpha v_h) - 2\alpha v_h = v_h(1-\alpha)$.

In the appendix, we prove that for $\alpha > \frac{1}{2}$, the provider always chooses to host each user on a different pool and will not benefit from serving both users from a single shared pool. Therefore, in equilibrium, the prices of the pools are set at αv_h and αv_l.

Expected Payoffs before knowing interruptions:

The provider's expected profit when each user minimizes and is hosted on a different pool is

$$E(\pi_{MM}) = (1-\gamma)^2\pi^{AA}_{MM} + \gamma(1-\gamma)\pi^{AI}_{MM} + \gamma(1-\gamma)\pi^{IA}_{MM} = 2(1-\gamma)\alpha(v_h + v_l)$$

The high value user's expected utility when both users minimize is

$$E(uHigh_{MM}) = (1-\gamma)^2 uHigh^{AA}_{MM} + \gamma(1-\gamma)uHigh^{AI}_{MM} = (1-\gamma)v_h(1-\alpha)$$

The low value user's expected utility when both users minimize is

$$E(uLow_{MM}) = (1-\gamma)^2 uLow^{AA}_{MM} + \gamma(1-\gamma)uLow^{IA}_{MM} = (1-\gamma)v_l(1-\alpha)$$

4.3.2 High-Value User Diversifies and Low-Value User Minimizes

In this case, one pool can host one of the high value user's instances, while the other pool hosts the other three instances. The prices are set by the provider after interruptions are known. Thus, the pricing depends on interruptions as follows.

(I) Both pools are available

Suppose both pools are available. Both of low value user's instances and one of the high value user's instances will come from the same pool. The remaining instance for the high value user will

come from the other pool. We denote the former pool, which hosts instances from both users, the shared pool. The pool with the single instance is denoted the exclusive pool. The provider sets the price of the exclusive pool to extract maximum rent from the high value user. Thus, we have $p_{exclusive} = v_h$.

The shared pool hosts three instances valued at αv_h, v_l, and αv_l. The price of the shared pool depends on which of the two values αv_h and v_l is bigger. The provider has three pricing options: 1) selling three instances at the price of αv_l, 2) selling two instances at the price of $Min[\alpha v_h, v_l]$, or 3) selling one instance at the price of $Max[\alpha v_h, v_l]$. If $\alpha v_h > v_l$, then the provider's optimal profit in the shared pool is $\pi^{AA}_{share} = Max[\alpha v_h, 2v_l, 3\alpha v_l]$. Otherwise, if $\alpha v_h < v_l$, then $\pi^{AA}_{share} = Max[v_l, 2\alpha v_h, 3\alpha v_l] = Max[2\alpha v_h, 3\alpha v_l]$, since $\alpha > 1/2$.

Assuming $\alpha v_h > v_l$, $Max[2v_l, 3\alpha v_l]$ is determined by α. If $\alpha > 2/3$ and $v_h < 3v_l$, we have $\pi^{AA}_{share} = Max[\alpha v_h, 2v_l, 3\alpha v_l] = 3\alpha v_l$ and $p^{AA}_{share} = \alpha v_l$. If $\alpha > 2/3$ and $v_h > 3v_l$, we have $p^{AA}_{share} = \alpha v_h$. Otherwise, if $\alpha < 2/3$, then $p^{AA}_{share} = v_l$ for $v_l < \alpha v_h < 2v_l$, and $p^{AA}_{share} = \alpha v_h$ for $2v_l < \alpha v_h$.

Assuming $\alpha v_h < v_l$, then $p^{AA}_{share} = \alpha v_l$ for $2v_h < 3v_l$, and $p^{AA}_{share} = \alpha v_h$ for $3v_l < 2v_h$. Again, if we assume $\alpha > 2/3$, since we have $\alpha v_h < v_l$ then $2v_h < 3v_l$ must hold and thus, $p^{AA}_{share} = \alpha v_l$.

Thus, if we assume $\alpha > 2/3$, the price of the shared pool and the provider's profit when both pools are available are shown below.

$$\begin{cases} p^{AA}_{share} = \alpha v_h \text{ and } \pi^{AA}_{DM} = v_h + \alpha v_h & \text{if } v_h > 3v_l \\ p^{AA}_{share} = \alpha v_l \text{ and } \pi^{AA}_{DM} = v_h + 3v_l & \text{if } v_h < 3v_l \end{cases} \tag{4}$$

Given the pool prices and the users' valuation of instances, we calculate the users' utility when both pools are available. As seen below, the only time a user gains surplus is when the high value user pays a price of for its second instance valued at in the shared pool.

$$\begin{cases} uHigh^{AA}_{DM} = 0 \text{ and } uLow^{AA}_{DM} = 0 & \text{if } v_h > 3v_l \\ uHigh^{AA}_{DM} = \alpha(v_h - v_l) \text{ and } uLow^{AA}_{DM} = v_l - \alpha v_l & \text{if } v_h < 3v_l \end{cases} \tag{5}$$

(II) The Exclusive pool is interrupted and the Shared pool is available

Next, suppose the exclusive pool is interrupted. Now, the high value user's instance in the shared pool is valued at v_h instead of αv_h, due to the interruption leaving this user with only one

instance. The provider has three pricing options: 1) selling three instances at the price of αv_l, 2) selling two instances at the price of v_l, or 3) selling one instance at the price of v_h.

Thus, the provider's optimal profit in the shared pool is $\pi^{IA}_{share} = Max[v_h, 2v_l, 3\alpha v_l]$. Assuming $\alpha > 2/3$, we know that selling three instances at the price of αv_l is better for the provider than selling two instances at the price of v_l, i.e. $2v_l < 3\alpha v_l$. For $v_h < 3\alpha v_l$, the provider prices the shared pool at $p^{IA}_{share} = \alpha v_l$ and earns a total profit of $\pi^{IA}_{DM} = 3\alpha v_l$. For $3\alpha v_l < v_h$, the provider prices the shared pool at $p^{IA}_{share} = v_h$ and earns a total profit of $\pi^{IA}_{DM} = v_h$. (Otherwise, if $\alpha < 2/3$, then $p^{IA}_{share} = v_l$ for $v_h < 2v_l$, and $p^{IA}_{share} = v_h$ for $2v_l < v_h$.) . In summary, for $\alpha > 2/3$, we have

$$\begin{cases} p^{IA}_{share} = v_h \text{ and } \pi^{IA}_{DM} = v_h & \text{if } v_h > 3\alpha v_l \\ p^{IA}_{share} = \alpha v_l \text{ and } \pi^{IA}_{DM} = 3\alpha v_l & \text{if } v_h < 3\alpha v_l \end{cases} \tag{6}$$

Given these prices, the users' utilities for is derived as follows.

$$\begin{cases} uHigh^{IA}_{DM} = 0 \text{ and } uLow^{IA}_{DM} = 0 & \text{if } v_h > 3\alpha v_l \\ uHigh^{IA}_{DM} = v_h - \alpha v_l \text{ and } uLow^{IA}_{DM} = v_l - \alpha v_l & \text{if } v_h < 3\alpha v_l \end{cases} \tag{7}$$

(III) The Shared pool is interrupted and the Exclusive pool is available

In this case the provider only sells one instance to the high value user. Thus, the exclusive pool is priced at $p_{exclusive} = v_h$ and the provider's profit is $\pi^{AI}_{DM} = v_h$. Neither user gains a surplus, i.e. $uHigh^{AI}_{DM} = uLow^{AI}_{DM} = 0$.

Expected Payoffs before knowing interruptions:

The provider's expected profit when the high value user diversifies and the low value user minimizes is The high value user's expected utility is The low value user's expected utility is

$$E(\pi_{DM}) = (1-\gamma)^2\pi^{AA}_{DM} + \gamma(1-\gamma)\pi^{AI}_{DM} + \gamma(1-\gamma)\pi^{IA}_{DM} = \begin{cases} (1-\gamma)^2(v_h + \alpha v_h) + 2\gamma(1-\gamma)v_h & \text{if } v_h > 3v_l \\ (1-\gamma)^2(v_h + 3\alpha v_l) + 2\gamma(1-\gamma)v_h & \text{if } 3\alpha v_l < v_h < 3v_l \\ (1-\gamma)(v_h + 3\alpha v_l) & \text{if } v_h < 3\alpha v_l \end{cases}$$

The high value user's expected utility is

$$E(uHigh_{DM}) = \begin{cases} 0 & \text{if } v_h > 3v_l \\ (1-\gamma)^2\alpha(v_h - v_l) & \text{if } 3\alpha v_l < v_h < 3v_l \\ (1-\gamma)^2\alpha(v_h - v_l) + \gamma(1-\gamma)(v_h - \alpha v_l) & \text{if } v_h < 3\alpha v_l \end{cases}$$

The low value user's expected utility is

$$E(uLow_{DM}) = \begin{cases} 0 & \text{if } v_h > 3v_l \\ (1-\gamma)^2 v_l(1-\alpha) & \text{if } 3\alpha v_l < v_h < 3v_l \\ (1-\gamma)v_l(1-\alpha) & \text{if } v_h < 3\alpha v_l \end{cases}$$

4.3.3 Both Users Diversify or Only Low-Value User Diversifies

When the low value user diversifies, this user will have the lowest valued instance in each pool, with the value being either v_l or αv_l. The price of each pool is either greater than or equal to the value for the lowest valued instance in that pool. Thus, there is no surplus for the low value user in either pool, i.e. $E(uLow_{MD}) = 0$ and $E(uLow_{DD}) = 0$. Therefore, diversification is always a dominated strategy for the low value user and in equilibrium the low value user minimizes.

4.4 Results

Comparing the high value user's expected utilities $E(uHigh_{MM})$ and $E(uHigh_{DM})$, we determine the user's optimal choice of diversification as a function of the likelihood of interruption.

Proposition 1 *Assuming the two users' value difference is not too high (i.e. $v_h < 3v_l$), the users' choice of diversification depends on the likelihood of interruption as follows:*

- *When the difference between the users' valuation of instances is low, only the high value user diversifies if the likelihood of interruption is high and both users minimize when the likelihood of interruption is low.*
- *When the difference between the users' valuation of instances is moderately high, only the high value user diversifies if the likelihood of interruption is low and both users minimize when the*

likelihood of interruption is high.

While common intuition suggests diversification is beneficial when interruptions are likely, this proposition shows that the opposite may in fact be true, if the provider's pricing decisions are taken into account. Given that diversification can allow users access to cloud resources, even if one of the pools becomes unavailable, it may seem that when chances of interruptions are high, users should opt to diversify. However, each user's choice of diversification can affect the price of the pools, depending on the value difference between the users. We show that the intuition that diversification happens for highly likely interruptions holds true only if the users are not highly different in their valuation of instances. Otherwise, for highly different users, diversification is chosen for lower probabilities of interruption.

When a user chooses to minimize, it can benefit from buying both instances at a low price equal to its valuation for its second instance. In other words, by pooling both instances in the same pool, the user can prevent the provider from price discriminating between the user's two instances. Other than this instance level price discrimination, there is also another form of price discrimination that can be employed by the provider, and that occurs across users. Namely, when both users choose to minimize, the provider can host each user's instances on a different pool. As a result, each user can receive a different price based on its valuation and the provider can successfully price discriminate between the two users. When one of the users chooses to diversify, the provider can no longer fully separate the two users' pricing options, and its ability to price discriminate across users is weakened. This is the main tradeoff of diversification for pricing of cloud pools: Minimization limits instance level price discrimination but allows for high user level price discrimination. Diversification strategies can alleviate this user level price discrimination, but at the cost of strengthening instance level price discrimination.

When the difference between the two users' valuations is low, the potential for user level price discrimination is not strong, since the two users are relatively similar and close in their valuation. Therefore, if both pools are likely to be available, users do not benefit greatly by limiting user level price discrimination through diversifying. Thus, in this case, the benefit of diversification for each user is dominated by the benefit of preventing instance level price discrimination and each user prefers to pool its instances and pay the minimum price of its second instance. However, this benefit only exists when there is no interruption in the pool allocated to the high value user. If

interruption is likely, there is a good chance that a diversified high value user can share its first and only instance with the low value user's two instances. Since the two users are not highly differentiated, the provider sets the price of the shared pool such that both users are able to purchase, even as the high value user's valuation of the instance in the shared pool raises with an interruption. This can prove beneficial for the high value user that gets to pay the price of the low value user's second instance. Thus under high chances of interruption, a diversified strategy becomes optimal for the high value user.

On the other hand, when there is higher difference between the two users' valuations, avoiding user level price discrimination can become much more attractive to the high value user. The high value user's choice of diversifying could allow the provider to price discriminate between this user's instances and selling one instance at the price of v_h. However, by pooling one of its instances with the low value user's instance, the high value user can benefit from the lower price of the pool allocated to the low value user, when there are no interruptions. If an interruption in the high value user's exclusive pool occurs, the high value user's only instance will be pooled with the low value user's instances. Since the two users' valuations are highly different, the provider benefits from setting a high price for this pool in order to extract all of the value from the high value user's first instance; a pricing strategy that was not optimal when this pool included the high value user's second instance under no interruption. Therefore, a diversified strategy can backfire for the high value user when interruptions are likely, as an interruption in the exclusive pool can significantly increase prices for the shared pool. In this case, the high value user diversifies, only when the likelihood of interruption is low.

Note that when value difference between the users is too high, i.e. $v_h > 3v_l$, both users always choose to minimize. This is because if the high value user diversifies, the provider prefers to capture the full value from both instances of the high value user. Thus, because of the high value user's extremely higher valuation the provider forgoes selling to the low value user and prices each pool to match the high value user's instance. Therefore, the high value user does not benefit from diversifying.

Figure 11, shows the choice of diversification depending on the value difference between the users and the likelihood of interruption. The shaded area, regions B and D, represents the area where the high value user diversifies and the low value user minimizes. In the remaining area,

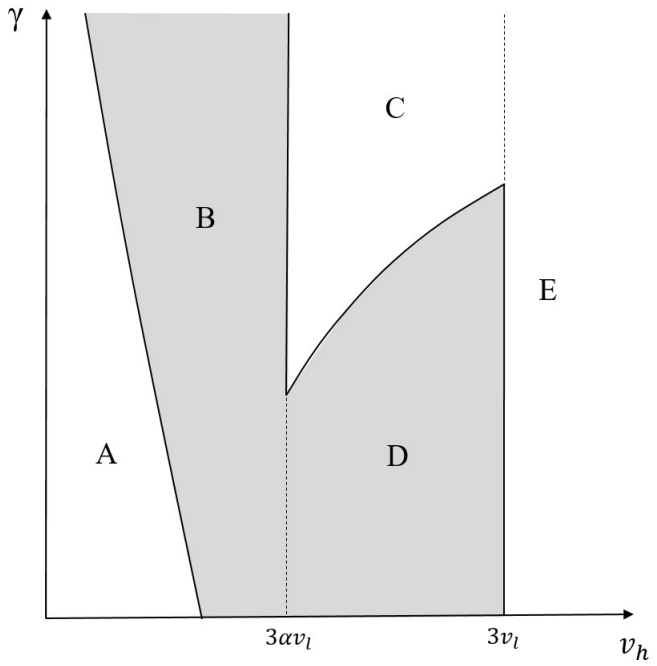

Figure 11: User's choice of diversifying cloud resources. (generated with $\alpha = 0.7$ and $v_l = 1$)

both firms choose to minimize. As shown in the figure, the effect of interruption likelihood on choice of diversity depends on the value difference between the users. For low value difference (i.e. $v_h < 3\alpha v_l$), diversification is chosen for high interruption likelihood, whereas for moderately high value difference (i.e. $3\alpha v_l < v_h < 3v_l$), diversification is chosen for low interruption likelihood. Finally, for extremely high value difference (i.e. $v_h > 3v_l$), diversification never occurs.

Proposition 2 *When the difference between the users' valuation of instances is moderately low and the high value user diversifies, an interruption in the exclusive pool increases the high value user's utility.*

Interestingly, this proposition shows interruptions may not always have a negative effect on the users' outcome. Specifically, when the two users are not highly differentiated, an interruption can benefit the high value user by lowering the average price per instance paid by this user. It may be intuitively expected that since each user can receive a positive net value for each of the two instances, removing one pool of instances should negatively impact every user that would be purchasing from that pool. However, we show that accounting for the price of the remaining pools and the valuation for the remaining instances can give us the opposite result.

Proposition 2 holds in region B of Figure 11, where the high value user diversifies and the value

difference between the users is low. In this region, when an interruption occurs in the exclusive pool where the high value user's first instance is hosted, the price of the shared pools remains the same, i.e. $p^{IA}_{share} = \alpha v_l$. Thus, as the exclusive is interrupted, the high value user is able to purchase its first and only instance from the shared instance, at a low price of αv_l. Note that without an interruption, the high-value user would pay v_h in the exclusive pool and αv_l in the shared pool. Thus the average price paid for each instance is higher without an interruption and an interruption can benefit the high value user.

Proposition 3 *The cloud provider's profit can be maximized at a non-zero level of value discount for the users' second instance, i.e.* $\alpha = \frac{v_h}{v_h+(1-\gamma)(v_h-v_l)}$*, when the likelihood of interruption is high (i.e.* $\gamma > \gamma^*$*) and the value difference between the two users is moderately high (i.e.* $3\alpha v_l < v_h < 3v_l$*).*

This proposition shows how the second instance value discount, $1 - \alpha$, affects the provider's profit. Given a fixed value of the first instances, the higher α is the higher both users value their second instances, resulting in a higher total value for the cloud provider's offering. Thus, one may expect that as the users' total value for instances gets higher, the cloud provider's profits should also increase. However, this proposition shows that may not be the case and the highest α does not necessarily result in the highest profit for the provider. We show that, despite lowering the users' overall value for the provider's offering, a medium level of value discount for the second instance can actually benefit the provider. This result is due to how α affects the users' decision to diversify or minimize.

When α is relatively low, the difference in value between first and second instances is high. Thus, the high value user greatly benefits from preventing instance level price discrimination and purchasing both instances at the price of its second instance. Therefore, the high value user chooses to minimize for low α. However, as α becomes higher, avoiding the instance level price discrimination becomes less attractive, since a user's two instances are valued closer to each other. Thus, for high α, the high value user chooses to diversify. Specifically, in the region specified in proposition 1(b), where the value difference between the two users is moderately high, the high value user switches from minimizing to diversifying at $\alpha = \alpha^* = \frac{v_h}{v_h+(1-\gamma)(v_h-v_l)}$ and diversifies for bigger α.

As shown in Figure 12, the provider's profit drops at $\alpha = \alpha^*$, where the high value user switches

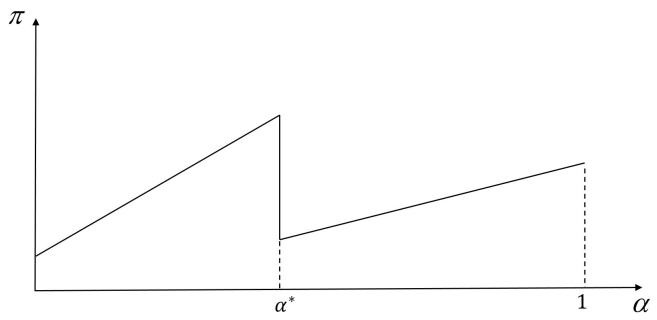

Figure 12: Provider's profit as a function of α(generated for $\gamma = 0.7$, $v_l = 1$, and $v_h = 2.5$)

to diversification. From this point, as α increases, the provider's profit also increases until α reaches 1. For high likelihood of interruption, γ, α^* is relatively high and close to 1. Thus, for high γ, the drop in the provider's profit occurs close to $\alpha = 1$ and increasing α from this point will not cause a big enough increase in the provider's profit to reach the maximum profit when both users minimized, even if $\alpha$is increased all the way to 1.

4.5 Discussion

In this essay, I analyzed spot pricing of computing resources in the cloud computing industry. Spot resources provide an affordable option for users willing to risk the uncertain availability of their computing resources. Cloud providers offer the option to diversify spot resources across multiple pools as a way for users to mitigate their losses in case one of the pools is interrupted. I study when it is optimal for users to use the diversification feature, depending on the likelihood that an interruption occurs. Intuitively, it may appear that users should opt for diversification when the likelihood of an interruption is high in order to avoid having all of their instances interrupted. However, I show that depending on the valuation of the users for computing resources, the opposite may in fact be true.

Considering the value difference between users of spot resources, I show that when users are relatively similar in their valuation of instances, the intuition that higher likelihood of interruption leads to diversification holds. However, when the users' valuation for spot instances are highly different, it is for lower likelihood of interruption that diversified allocation of resources is chosen. The reason behind this finding is that the provider's pricing decision is a function of interruption occurrences as well as users' diversification choices. When users diversify, the provider can no

longer set different prices for each user and must choose a single price for the pool shared by both users. Thus, if the users' valuations are highly different, the provider finds it optimal to only sell to the high value user, if only the shared pool is available due to an interruption in the other pool. This means an interruption can result in a significantly higher price in the remaining shared pool if diversification is chosen. Therefore, diversification is not an optimal strategy for users when their value difference is high.

On the other hand, when the users' valuation of instances are similar and they share a pool of resources, the provider benefits from selling to both users, even if other pools are interrupted. This means interruptions will not increase the price of the shared pool and the high value user can benefit from paying the same low price as the low value user. Specifically for the high value user, we show that when this user diversifies and an interruption occurs in the exclusive pool, the high value user's payoff is increased. This is because as a result of the interruption, the high value user only purchases one instance in the shared pool, receiving a high value at a low price. Thus, for low value differentiation between the users and high likelihood of interruption, the high value user diversifies and actually benefits from an interruption.

The findings from this essay show when it is optimal for users to choose diversification when using spot resources. We also studied the effect of diversification on the provider's profits, depending on the users' valuation for multiple instances. Future research in this area can look at the users' choice between using spot resources and on-demand resources and when it is optimal to use each pricing option.

5 Conclusion

In this dissertation, I study how cloud computing technology changes marketing practices. Cloud computing has had a significant impact on the way firms offer web-based services and market their services. In particular, cloud services have revolutionized capacity choices of web-based firms by offering on-demand capacity using remote computation. This also brings new approaches for pricing computational resources. Specifically, I study two major features of the cloud computing industry: autoscaling and spot pricing. For each of these two feature, I identify implications for marketing decisions made by cloud computing users as well as cloud computing service providers.

The autoscaling feature of cloud computing allows firms to forgo having to invest in capacity prior to demand realization. I show the many applications of this feature on entrepreneurs entering new markets. Intuitively, one might expect the introduction of autoscaling to increase the number of new entrants in the market, since it can reduce the risk of entry by eliminating capacity decisions. However, we show that this may not always be the case and autoscaling can decrease or increase the number of entrants depending on the cost of entry and the probability of each firm's success. Autoscaling always increases the range of entry cost for which at least one firm enters the market, consistent with general intuition. However, we find that autoscaling can counterintuitively decrease the range of entry costs for which both firms enter the market when the probability of success is high for each firm. The ability to automatically scale up to meet demand serves as a double edged sword since it gives firms license to aggressively pursue market share. This causes firms to anticipate higher competition when the probability of both firms offering high value is high enough, which in turn results in less likelihood of both firms entering the market.

In the context of spot pricing in cloud computing, we studied the implications of diversification strategies. The ability to diversify allows users to use resources from a multitude of instance pools, reducing the chance that all of their instances become unavailable due to interruptions. This research studies the conditions under which users should adopt a diversified allocation of cloud resources. We show that when the users are similar in their valuation of cloud resources, diversification is optimal when the likelihood of an interruption is high. However, when the users are highly differentiated in their valuation, the opposite holds and diversification is chosen when interruption likelihood is low. The findings identify the valuation difference among users, the

valuation difference between first and second instances for each user, and the likelihood of an interruption as key factors determining when a diversification strategy is optimal for users.

This research has clear implications for marketing managers. The first essay guides managers of startups with their response to the introduction of autoscaling and cloud computing. The research shows if the probability of success is moderately high, autoscaling leads to a situation where both startups adopt autoscaling, but they would have been better off if both had avoided it. Our findings suggest that in industries where cost of entry is relatively low, the introduction of the autoscaling feature in cloud computing can reduce the profit of new entrants to the market even though it decreases their downside risk of failure. The second essay guides users to choose the proper allocation strategies for their spot resources in order to optimize their usage of the spot market without incurring large losses due to potential interruptions. The findings also inform managers on whether or not an interruption is beneficial to a user, depending on the allocation strategy chosen.

Future research on cloud computing can consider other unique features offered by cloud providers. For instance, another unique offering of the cloud providers is the hybrid cloud, which allows users to host applications partially on cloud servers and partially on their in-house servers, while having the ability to communicate between the two infrastructures. Orchestration between on premises and cloud services is a key component of the hybrid cloud, which required additional technology so that the private infrastructure is able to run cloud services (Sanders, 2014). Adoption of the hybrid cloud has been gaining momentum, with an estimated 55% of enterprises using hybrid cloud (Walden, 2015). Future studies can explore the implications of the hybrid cloud and inform users on the optimal conditions under which a hybrid solution should be chosen.

This research is one of the first to study the implications of cloud computing for marketing decisions. I believe, the future of the field of marketing can be largely impacted by cloud computing technology and plan to provide the basis for extensive research on this topic with this dissertation.

References

[1] Aksin, Z., M. Armony, and V. Mehrotra. 2007. "The modern call center: A multi-disciplinary perspective on operations management research." *Production and Operations Management* 16(6): 665-688.

[2] Aksin, Z., F. De Vericourt, and F. Karaesmen. 2008. "Call center outsourcing contract analysis and choice." *Management Science* 54(2): 354-368.

[3] Anupindi, R., and R. Akella. 1993. "Diversification under supply uncertainty." *Management Science* 39(8): 944-963.

[4] Anupindi, R. and L. Jiang. 2008. "Capacity Investment Under Postponement Strategies, Market Competition, and Demand Uncertainty." *Management Science* 54(11): 1876-1890.

[5] Armbrust, M., A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica. 2009. "Above the clouds: A Berkeley View of Cloud Computing." *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS* 28, 13.

[6] Babich, V., A. Burnetas, and P. Ritchken. 2007. "Competition and diversification effects in supply chains with supplier default risk." *Manufacturing and Service Operations Management* 9(2): 123-146.

[7] Biyalogorsky, E., and O. Koenigsberg. 2014. "The Design and Introduction of Product Lines When Consumer Valuations are Uncertain." *Production and Operations Management* 23(9): 1539-1548.

[8] Bort, C. 2015. "Google Has Nabbed Some Huge Customers for its Most Important New Business Cloud Computing." *http://www.businessinsider.com/google-names-huge-companies-using-its-cloud-2015-6*, accessed September 2016.

[9] Buehler, S. and J. Haucap. 2006. "Strategic Outsourcing Revisited." *Journal of Economic Behavior and Organization* 61(3) 325-338.

[10] Cachon, G. and P. Harker. 2002. "Competition and outsourcing with scale economies." *Management Science* 48(10): 1314-1333.

[11] Che, H., C. Narasimhan, and V. Padmanabhan. 2010. "Leveraging Uncertainty Through Backorder." *Quantitative Marketing and Economics* 8(3): 365-392.

[12] Chen, P. and S. Wu. 2013. "the Impact and Implications of On-Demand Services on Market Structure." *Information Systems Research* 24(3): 750-767.

[13] Chen, Y. and G. Iyer. 2002. "Consumer Addressability and Customized Pricing." *Marketing Science* 21(2): 197-208.

[14] Columbus, L., 2015. "Roundup of Cloud Computing Forecasts and Market Estimates Q3 Update, 2015." *www.forbes.com/sites/louiscolumbus/2015/09/27/roundup-of-cloud-computing-forecasts-and-market-estimates-q3-update-2015/*, accessed September 2016.

[15] Dadda, M., N. Petruzzi, and L. Schwarz. 2007. "A newsvendor's procurement problem when suppliers are unreliable." *Manufacturing and Service Operations Management* 9(1): 9-32.

[16] Desai, P., O. Koenigsberg, and D. Purohit. 2007. "The Role of Production Lead Time and Demand Uncertainty in Marketing Durable Goods." *Management Science* 53(1): 150-158.

[17] Desai, P., O. Koenigsberg, and D. Purohit. 2010. "Forward Buying by Retailers." *Journal of Marketing Research* 47(1): 90-102.

[18] Daughety, A. 1990. "Beneficial Concentration." *American Economic Review* 80(5): 1231-1237.

[19] Federgruen, A. and N. Yang. 2009. "Optimal supply diversification under general supply risks." *Operations Research* 57(6): 1451-1468.

[20] Feng, Q. and L. Lu. 2011. "The strategic perils of low cost outsourcing." *Management Science* 58(6): 1196-1210.

[21] Feng, Q. and R. Shi. 2012. "Sourcing from multiple suppliers for price-dependent demands." *Production and Operations Management* 21(3): 547-563.

[22] Ferguson, M. and O. Koenigsberg. 2007. "How Should a Firm Manage Deteriorating Inventory." *Production and Operations Management* 16(3): 306-321.

[23] Flood, G. 2013. "Gartner Tells Outsourcers: Embrace Cloud Or Die." *http://www.informationweek.com/cloud/infrastructure-as-a-service/gartner-tells-outsourcers-embrace-cloud-or-die/d/d-id/1110991*, accessed September 2016.

[24] Gans, N. and Y. Zhou. 2003. "A call-routing problem with service-level constraints." *Operations Research* 51(2): 255-271.

[25] Gaudin, S. 2015. "AWS Assures Customers it's Atop the Cloud." *http://www.computerworld.com/article/2990119/cloud-computing/aws-assures-customers-theyre-atop-the-cloud.html*, accessed September 2016.

[26] Griffith, E. 2014. "Why Startups Fail, According to Their Founders." *Forbes* Sep, 25.

[27] Grossman, G. and E. Helpman. 2002. "Integration versus outsourcing in industry equilibrium." *The Quarterly Journal of Economics* 117(1): 85-120.

[28] Gupta, P., A. Seetharaman, and J. R. Raj. 2013. "The Usage and Adoption of Cloud Computing by Small and Medium Businesses." *International Journal of Information Management* 33(5): 861-874.

[29] Hasija, S., E. Pinker, and R. Shumsky. 2008. "Call center outsourcing contracts under information asymmetry." *Management Science* 54(4): 793-807.

[30] Hassani, S. 2009. *Mathematical Methods for Students of Physics and Related Fields* (2nd ed.). New York, NY: Springer. Chapter 5.

[31] He, B., H. Huang, and K. Yuan. 2016. "Managing supply disruption through procurement strategy and price competition." *International Journal of Production Research* 54(7): 1980-1999.

[32] Iyer, G., D. Soberman, and J. Villas-Boas. 2005. "The Targeting of Advertising." *Marketing Science* 24(3): 461-476.

[33] Joshi, Y., D. Reibstein, and J. Zhang. 2009. "Optimal Entry Timing in Markets with Social Influence." *Management Science* 55(6): 926-939.

[34] Kreps, D. and J. Scheinkman. 1983. "Quantity Precommitments and Bertrand Competition Yield Cournot Outcomes." *The Bell Journal of Economics* 14(2): 326-337.

[35] Leavitt, N. 2009. "Is Cloud Computing Really Ready For Prime Time?" *Computer* (1): 15-20.

[36] Lilien, G. and E. Yoon. 1990. "The Timing of Competitive Market Entry: An Exploratory Study of New Industrial Products." *Management Science* 36(5): 568-585.

[37] Marston, S., Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi. 2011. "Cloud Computing: The Business Perspective." *Decision Support Systems* 51(1): 176-189.

[38] Mell, P. and T. Grance. 2011. "The NIST definition of cloud computing." *National institute of standards and technology.*

[39] Milgrom, P. and J. Roberts. 1982. "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis." *Econometrica* 50(2): 443-459.

[40] Narasimhan, C. and J. Zhang. 2000. "Market Entry Strategy Under Firm Heterogeneity and Asymmetric Payoffs." *Marketing Science* 19(4): 313-327.

[41] Narasimhan, C. 1988. "Competitive Promotional Strategies." *Journal of Business* 61(4): 427-449.

[42] Nasser, S. and D. Turcic. 2015. "To Commit or Not to Commit: Revisiting Quantity vs. Price Competition in a Differentiated Industry." *Management Science* 62(6): 1719-1733..

[43] Nickelsburg, M. 2016. "Startup Spotlight: BeFunky makes online photo editing and graphic design simple." *http://www.geekwire.com/2016/befunky/*, accessed September 2016.

[44] Ofek, E. and O. Turut. 2013. "Vaporware, Suddenware, and Trueware: New Product Preannouncements Under Market Uncertainty." *Marketing Science* 32(2): 342-355.

[45] Parlar, M. and D. Berkin. 1991. "Future supply uncertainty in EOQ models." *Naval Research Logistics* 38(1): 107-121.

[46] Poepsel, M. 2008. "Customers Are Won or Lost in One Second, Finds New Aberdeen Report" *http://www.prnewswire.com/news-releases/customers-are-won-or-lost-in-one-second-finds-new-aberdeen-report-65399152.html*, accessed September 2016.

[47] Ray, T. 2013. "Salesforce, Google, Amazon Cloud Winners, Says Piper; Microsoft Straddles the Line." *blogs.barrons.com* Oct, 17.

[48] Ren, Z. and Y. Zhou. 2008. "Call center outsourcing: Coordinating staffing level and service quality." *Management Science* 54(2): 369-383.

[49] Reynolds, S. and B. Wilson. 2000. "Bertrand-Edgeworth Competition, Demand Uncertainty, and Asymmetric Outcomes." *Journal of Economic Theory* 92: 122-141.

[50] Shulman, J. 2014. "Product Diversion to a Direct Competitor." *Marketing Science* 33(3): 422-436.

[51] Shy, O. and R. Stenbacka. 2003. "Strategic outsourcing." *Journal of Economic Behavior and Organization* 50(2): 203-224.

[52] Shy, O. and R. Stenbacka. 2005. "Partial outsourcing, monitoring cost, and market structure." *Canadian Journal of Economics* 38(4): 1173-1190.

[53] Spence, M. 1977. "Entry, Capacity, Investment and Oligopolistic Pricing." *The Bell Journal of Economics* 8(2): 534-544.

[54] Sultan, N. A. 2011. "Reaching for the Cloud: How SMEs Can Manage." *International journal of information management* 31(3):272-278.

[55] Swinney, R., G. Cachon., and S. Netessine. 2011. "Capacity Investment Timing by Start-ups and Established Firms in New Markets." *Management Science* 57(4): 763-777.

[56] Tirole, J. 1988. *The Theory of Industrial Organization.* Cambridge, MA: MIT Press.

[57] Van Mieghem, J. 1999. "Coordinating investment, production, and subcontracting." *Management Science* 45(7): 954-971.

[58] Van Mieghem, J. and M. Dada. 1999. "Price Versus Production Postponement: Capacity and Competition." *Management Science* 45(12):1631-1649.

[59] Walker, E. 2009. "The Real Cost of a CPU Hour." *Computer* (4): 35-41.

[60] Weaver, L. 2015. "Why GE is moving nearly 9,000 apps to the public cloud." *https://enterprisersproject.com/article/2015/10/why-ge-s-moving-nearly-9000-apps-public-cloud*, accessed September 2016.

[61] Work, S. 2011. "How Loading Time Affects Your Bottom Line." *https://blog.kissmetrics.com/loading-time/*, accessed September 2016.

[62] Wu, X., and F. Zhang. 2014. "Home or overseas? An analysis of sourcing strategies under competition." *Management Science* 60(5): 1223-1240.

[63] Yang, H., and M. Tate. 2012. "A Descriptive Literature Review and Classification of Cloud Computing Research." *Communications of the Association for Information Systems* 31(2): 35-60.

[64] Yano, C., and H. Lee. 1995. "Lot sizing with random yields: A review." *Operations Research* 43(2): 311-334.

[65] Zhang, K. and Z. Katona. 2012. "Contextual Advertising." *Marketing Science* 31(6): 980-994.

[66] Zhou, B., C. Mela, and W. Amaldoss. 2015. "Do Firms Endowed with Greater Strategic Capability Earn Higher Profits?" *Journal of Marketing Research* 52(3): 325-336.

A Technical Appendix

Proof of Proposition 1 in Essay 1

If $\gamma < c/V$, then expected profit is strictly decreasing in capacity for any capacity chosen by the competition. Therefore, each firm optimally chooses zero capacity.

If $\gamma > c/V$, then for any $k_2 < 1-\alpha$, expected profit is increasing in k_1 for any $k_1 < 1-k_2$. Therefore, we can rule out any $k_1 + k_2 < 1$. First consider $\gamma(1-\gamma)V - c < 0$. Suppose there were a symmetric equilibrium in which $k_1 + k_2 > 1$. By definition, this implies that $k_1 > 1/2$ which means Firm 1 earns greater expected profit by deviating downward. Therefore, the only potential symmetric equilibrium requires $k_1 + k_2 = 1$. If Firm 1 deviates upward, its expected change in profit is $\gamma(1-\gamma)V - c < 0$.

If $\gamma(1-\gamma)V - c > 0$, then $k_1 + k_2 = 1$ is no longer an equilibrium because either firm can profitably deviate to harvest the potential of monopoly power. We can therefore focus our attention on $k_1 + k_2 > 1$. If $k_2 > k_1$, then Firm 2's expected profit is strictly increasing in k_2 until $k_2 = 1-\alpha$ and is strictly decreasing for any $k_2 > 1-\alpha$. If $k_2 > k_1$, Firm 1's expected profit for any $k_1 < 1-\alpha$ is given by $\gamma(1-\gamma)V + \gamma^2\frac{(1-k_1)k_1V}{k_2} - k_1c$ which is concave in k_1 and maximized at $k_1 = \frac{\gamma V(1-\alpha(1-\gamma))-c(1-\alpha)}{2\gamma^2V}$. Note this value of k_1 is in fact less than $1-\alpha$ if and only if $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$. We verify that Firm 2 cannot benefit from a global deviation undercutting k_1 at this level. Supposing that Firm 2 deviates to a lower capacity than Firm 1, its profit would become $k_2\left(-c+\gamma V\left(1-\gamma-\frac{2\gamma^3(-1+k_2)V}{c(-1+\alpha)+\gamma V(1+(-1+\gamma)\alpha)}\right)\right)$, which is maximized at

$$\tilde{k}_2 = -\frac{c^2(-1+\alpha)+c\gamma V(2+2\alpha(-1+\gamma)-\gamma)+V^2\gamma^2\left(-1+\alpha+\gamma-2\alpha\gamma+(-2+\alpha)\gamma^2\right)}{4V^2\gamma^4}.$$

It is easily shown that any values of c that allow for $\tilde{k}_2 < k_1$ will result in Firm 2's profit at $\tilde{k}_2$ (i.e., $\frac{\left(c^2(-1+\alpha)+c\gamma V(2-\gamma+2(-1+\gamma)\alpha)+\gamma^2V^2\left(-1+\gamma+\gamma^2(-2+\alpha)+\alpha-2\gamma\alpha\right)\right)^2}{8\gamma^4V^2(c(-1+\alpha)+\gamma V(1+(-1+\gamma)\alpha))}$) to be less than Firm 2's profit at our equilibrium.

If $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$ then neither firm benefits from deviating from $k_j = 1-\alpha$.

Proof of Lemma 1 in Essay 1

A firm will enter if its expected profit is greater than its entry cost. If and only if $F < (\gamma V - c)(1-\alpha)$, then a firm's best response to its competitor not entering the market is to enter the market. Supposing one firm enters the market, the remaining firm's best response to its competitor's entry is to also enter the market provided the expected profit earned when competing is greater than the cost of entry. The anticipated payoffs associated with being one of two firms entering are reported in Proposition 1 and the conditions on F are presented in the Lemma.

Analysis of Pricing Subgame When Only One Firm Uses Autoscaling in Essay 1

We start by showing that this game does not have a pure strategy equilibrium. Assume for sake of contradiction that the firms use prices p_1 and p_2 in a pure strategy equilibrium. If $p_1 \neq p_2$, then the firm with a lower price can benefit from deviating by increasing its price to $\frac{p_1+p_2}{2}$. If $p_1 = p_2$, then Firm 2 can benefit from deviating by decreasing its price to $p_2 - \varepsilon$, for sufficiently small ε, to acquire all consumers in Segment 3. Therefore, a pure strategy equilibrium cannot exist.

Next, we find a mixed strategy equilibrium for this game. Provided $k_1 \leq 1 - \alpha$, Firm 2 can choose to *attack* with a price that clears its capacity or *retreat* with a price V that harvests the value from the $1 - k_1$ consumers that Firm 1 cannot serve due to its capacity constraint. Let z'' be the price at which Firm 2 is indifferent between *attacking* to sell to $1 - \alpha$ consumers at price z'' and *retreating* to sell to $1 - k_1$ consumers at price V. We have $(1 - k_1)(V - c) = (1 - \alpha)(z'' - c)$ which gives us $z'' = \frac{\alpha c - c k_1 + (k_1 - 1)V}{\alpha - 1}$. In equilibrium, both firms use a mixed strategy with prices ranging from z'' to V. Suppose that $H_i(.)$ is the cumulative distribution function used by Firm i. The profit of Firm 1 earned by setting price x is

$$\pi_1(x) = H_2(x)\alpha x + (1 - H_2(x))k_1 x - k_1 c$$

Using equilibrium conditions, we know that the derivative of this function with respect to x must

be zero for $x \in (z'', V)$. By solving the differential equation we get

$$H_2(x) = \begin{cases} 0 & \text{if } x < z'' \\ \frac{k_1(-ck_1+(-1+k_1)V+\alpha(c-x)+x)}{(-1+\alpha)(\alpha-k_1)x} & \text{if } z'' \le x < V \\ 1 & \text{if } x \ge V \end{cases}$$

Similarly, the profit of Firm 2 earned by setting price x is

$$\pi_2(x) = H_1(x)(1-k_1)(x-c) + (1-H_1(x))(1-\alpha)(x-c)$$

By setting the derivative with respect to x to zero for $x \in (z'', V)$ and solving the differential equation, we get

$$H_1(x) = \begin{cases} 0 & \text{if } x < z'' \\ \frac{-ck_1+(-1+k_1)V+\alpha(c-x)+x}{(k_1-\alpha)(x-c)} & \text{if } z'' \le x < V \\ 1 & \text{if } x \ge V \end{cases}$$

Proof of Lemma 2 in Essay 1

We first show that both firms not using autoscaling cannot be an equilibrium. The expected profit of using autoscaling, when the opponent uses capacity k, is

$$\pi_{AN} = \gamma^2(1-k)(V-c) + \gamma(1-\gamma)(1-\alpha)(V-c).$$

The expected profit of using fixed capacity k', when the opponent uses capacity k, according to the proof of Proposition 1, is

$$\pi_{NN} = \gamma(1-\gamma)k'(V-c) - (1-\gamma)k'c + \gamma^2 \times \begin{cases} \frac{(1-k')k'V}{k} - k'c & \text{if } k' \le k \\ \min(k', 1-k)V - k'c & \text{if } k' > k \end{cases}$$

First, note that if $k \le \frac{1}{2}$, then $\begin{cases} \frac{(1-k')k'V}{k} - k'c & \text{if } k' \le k \\ \min(k', 1-k)V - k'c & \text{if } k' > k \end{cases}$ is optimized at $k' = 1-k$. Therefore, we have $\pi_{AN} > \pi_{NN}$, i.e., deviating to autoscaling is profitable when the opponent uses

fixed capacity $k \leq \frac{1}{2}$. Thus, for an equilibrium in which neither firm uses autoscaling to exist, both firms must be using fixed capacity larger than half.

We know from Proposition 1 that the only possible equilibrium in which both firms use fixed capacity larger than $\frac{1}{2}$ (i.e., that satisfies best response equilibrium condition) is the one in which they both set their capacities to $1-\alpha$. But in this case, we have $\pi_{NN} = (1-\alpha)(V(\gamma(1-\gamma)) - c) + \gamma^2 V\alpha$, which is strictly less than π_{AN} for feasible k. Therefore, deviating to autoscaling is strictly profitable for each firm. Therefore, both firms using fixed capacities cannot be an equilibrium.

Given the fact that at least one firm uses autoscaling in equilibrium, we examine whether both firms use autoscaling or only one. If Firm 2 adopts autoscaling, Firm 1's best response is to adopt autoscaling if and only if

$$(V-c)(\gamma^2\alpha + \gamma(1-\gamma)(1-\alpha)) \geq \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(1-\alpha)\gamma^2(V-c)}$$

where the right hand side of the inequality is Firm 1's profit if it chooses the best possible capacity k, given by $k_1^* = \frac{c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha-1-\alpha\gamma)}{2\gamma^2(c-V)}$, to Firm 2's autoscaling decision. By rearranging this inequality for c/V, we get the condition in Lemma 2. Finally, we can confirm that when the above inequality does not hold and $k_1^* = \frac{c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha-1-\alpha\gamma)}{2\gamma^2(c-V)}$, Firm 2 does not benefit from deviating to fixed capacity. Therefore, when the above inequality does not hold, Firm 1 uses fixed capacity k_1^* and Firm 2 uses autoscaling in equilibrium.

If only one firm enters, this firm will enjoy monopoly power over $(1-\alpha)$ consumers with probability γ. Using autoscaling, the firm's profit equals to $\gamma(V-c)(1-\alpha)$, which is strictly larger than $(V\gamma - c)(1-\alpha)$, the profit earned without autoscaling, for any $\gamma > 0$.

Proof of Proposition 2 in Essay 1

We start by calculating the average prices without autoscaling when $\gamma > c/V$, $\gamma(1-\gamma)V - c > 0$, and $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$. Based on Proposition 1, both firms set their capacities equal to $1-\alpha$. The cumulative distribution function of prices set by each firm is $F(x) = \frac{(1-\alpha)x+V\alpha}{(1-2\alpha)x}$. Thus, the probability density function equals $f(x) = \frac{V\alpha}{x^2(1-2\alpha)}$ and the average price of each firm is $\bar{p}_{NN} = \int_z^V f(x)x\,dx = \frac{\alpha V \log(\frac{1-\alpha}{\alpha})}{1-2\alpha}$. With both firms using autoscaling, the probability density function equals $g(x) = \frac{\alpha(V-c)}{(1-2\alpha)(x-c)^2}$ and the average price of each firm becomes $\bar{p}_{AA} = \int_{z'}^V g(x)x\,dx =$

$\frac{(1-2\alpha)c+\alpha\log\left(\frac{1-\alpha}{\alpha}\right)(V-c)}{1-2\alpha}$. Thus we have $\bar{p}_{AA}-\bar{p}_{NN}=c(1-\frac{\alpha\log(\frac{1-\alpha}{\alpha})}{1-2\alpha})$. For all $s>1$, we know that $\log(s)<s-1$. Allowing $s=\frac{1-\alpha}{\alpha}$, we show that $\log(\frac{1-\alpha}{\alpha})<\frac{1-2\alpha}{\alpha}$, thus $\bar{p}_{AA}>\bar{p}_{NN}$. When only one firm (Firm 2) uses autoscaling and the other (Firm 1) chooses capacity k_1, the probability density function of Firm 1's price equals $h_1(x)=-\frac{(k_1-1)(c-V)}{(\alpha-k_1)(c-x)^2}$ and its average price is $\bar{p}_{NA}=\int_{z''}^{V}h_1(x)x\,dx=c+\frac{(1-k_1)(V-c)\log\left(\frac{1-\alpha}{1-k_1}\right)}{k_1-\alpha}$. We show $\frac{\partial(\bar{p}_{NA})}{\partial k_1}=-\frac{(c-V)\left((\alpha-1)\log\left(\frac{\alpha-1}{k_1-1}\right)-\alpha+k_1\right)}{(\alpha-k_1)^2}$ is negative. For all $s>0$, we know that $s\log(s)>s-1$. Allowing $s=\frac{1-\alpha}{1-k_1}$, we have $\frac{\alpha-k_1}{k_1-1}<\frac{(1-\alpha)Log((1-\alpha)/(1-k_1))}{(1-k_1)}$ and $\frac{\partial(\bar{p}_{NA})}{\partial k_1}<0$. Therefore, the minimum of $\bar{p}_{NA}$ occurs at $k_1=1-\alpha$ and is equal to $\bar{p}_{AA}$. Thus, $\bar{p}_{NA}\geq\bar{p}_{AA}>\bar{p}_{NN}$.

For Firm 2, which uses autoscaling, the probability density function equals

$$h_2(x)=\frac{k_1(-\alpha c+ck_1-k_1V+V)}{(\alpha-1)x^2(\alpha-k_1)}+\left(1-\frac{k_1(c-V)}{(\alpha-1)V}\right)\delta(x-V)$$

and average price equals

$$\bar{p}_{AN}=\int_{z''}^{V}h_2(x)x\,dx=\frac{(\alpha-k_1)(V(\alpha+k_1-1)-ck_1)+k_1(-\alpha c+ck_1-k_1V+V)\log\left(\frac{(1-\alpha)V}{-\alpha c+ck_1-k_1V+V}\right)}{(\alpha-1)(\alpha-k_1)}$$

We show $\frac{\partial(\bar{p}_{AN})}{\partial k_1}=\frac{\left(V\left(-2\alpha k_1+\alpha+k_1^2\right)-c(\alpha-k_1)^2\right)\log\left(\frac{(1-\alpha)V}{c(k_1-\alpha)+V(1-k_1)}\right)-\alpha(\alpha-k_1)(c-V)}{(\alpha-1)(\alpha-k_1)^2}$ is negative. For all $s>0$, we know that $s\log(s)>s-1$. Allowing $s=\frac{(1-\alpha)V}{c(k_1-\alpha)+V(1-k_1)}$, we find $\frac{\partial(\bar{p}_{AN})}{\partial k_1}<\frac{(\alpha-k_1)(c-V)^2}{(\alpha-1)^2V}<0$. Therefore, the minimum of $\bar{p}_{AN}$ occurs at $k_1=1-\alpha$ and we have $Min[\bar{p}_{AN}]-\bar{p}_{NN}=c\log\left(\frac{(1-\alpha)V}{-2\alpha c+\alpha V+c}\right)+\frac{\alpha V\log\left(\frac{-2\alpha c+\alpha V+c}{\alpha V}\right)}{2\alpha-1}+c$. Since $\log\left(\frac{-2\alpha c+\alpha V+c}{\alpha V}\right)<\frac{-2\alpha c+\alpha V+c}{\alpha V}-1$, we know $Min[\bar{p}_{AN}]-\bar{p}_{NN}>\frac{(\alpha V)\left(\frac{-2\alpha c+\alpha V+c}{\alpha V}-1\right)}{2\alpha-1}+c\left(\log\left(\frac{(1-\alpha)V}{-2\alpha c+\alpha V+c}\right)+1\right)=c\log\left(\frac{(1-\alpha)V}{-2\alpha c+\alpha V+c}\right)>0$. Thus, $\bar{p}_{AN}>\bar{p}_{NN}$.

When $\gamma>c/V$ and $\gamma(1-\gamma)V-c<0$, we know from Proposition 1 that without autoscaling, $k_1+k_2=1$ and both firms set their price equal to V. As we showed, $\log(\frac{1-\alpha}{\alpha})<\frac{1-2\alpha}{\alpha}$. Thus, $\bar{p}_{AA}=\frac{(1-2\alpha)c+\alpha\log\left(\frac{1-\alpha}{\alpha}\right)(V-c)}{1-2\alpha}<V$. Since $\log(\frac{1-\alpha}{1-k_1})<\frac{1-\alpha}{1-k_1}-1=\frac{k_1-\alpha}{1-k_1}$, we have $\bar{p}_{NA}<V$. Finally, we have $\log\left(\frac{(1-\alpha)V}{-\alpha c+ck_1-k_1V+V}\right)<\frac{(1-\alpha)V}{-\alpha c+ck_1-k_1V+V}-1$ and therefore, $\bar{p}_{AN}<V$.

Proof of Lemma 3 in Essay 1

Given the best capacity response to autoscaling, $k_1^* = \frac{V\gamma(1-\alpha+\alpha\gamma)-c\left(1+\alpha\left(-1+\gamma^2\right)\right)}{2(V-c)\gamma^2}$, the expected profit for Firm 1 choosing not to autoscale, assuming that Firm 2 uses autoscaling, is

$$\pi_{NA} = \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(1-\alpha)\gamma^2(V-c)}$$

and the expected profit for Firm 2 choosing to autoscale, assuming that Firm 1 uses fixed capacity k_1^*, is

$$\pi_{AN} = \frac{(\gamma V(\alpha(\gamma-1)+1)-c(\alpha(\gamma-1)^2+2\gamma-1))}{2}.$$

It is straightforward to show $\pi_{NA} < \pi_{AN}$. Thus, if the subgame equilibrium involves only one firm choosing autoscaling, both firms will enter if $F < \pi_{NA}$.

Both firms will choose autoscaling if $\pi_{NA} < \pi_{AA}$. As such, both firms will enter if $F < Max[\pi_{NA}, \pi_{AA}]$.

Proof of Proposition 3 in Essay 1

With autoscaling, at least one firm enters the market if and only if $F < F_A \equiv \gamma(V-c)(1-\alpha)$ whereas without autoscaling at least one firm enters if and only if $F < F_N \equiv (\gamma V - c)(1-\alpha)$.

Proof of Proposition 4 in Essay 1

We compare the cut-offs on F such that both firms enter. Let F_{AA} denote the cut-off for two firms to enter with autoscaling, F_{NA} denote the cut-off for two firms to enter when only one firm uses autoscaling, and F_{NN} denote the cut-off for two firms to enter when autoscaling is not available.

First consider the two cases in which $\gamma > c/V$ and $\gamma(1-\gamma)V - c > 0$. If $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$, then $0 < c < V$ requires $\gamma < (1-\alpha)/(2-3\alpha)$. Therefore, based on Lemma 2, both firms will use autoscaling if autoscaling is available. Thus, $F_{AA} - F_{NN} = c(1-\alpha+\gamma(1-\gamma-\alpha+2\alpha\gamma))$ which is decreasing in α and positive at $\alpha = 1/2$ and therefore positive for all $\alpha < 1/2$. If $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$, the derivative of $F_{AA} - F_{NN}$ with respect to V is $\frac{c^2(1-\alpha)^2-\gamma^2V^2(1-\alpha+\gamma(3\alpha-2))^2}{4\gamma^2V^2(1-\alpha)}$, which is positive for all $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$. At $V = \frac{c}{\gamma}$, we have $F_{AA} - F_{NN} = \frac{1}{4}c\left(4(1-2\alpha)\gamma^2 + \frac{(\alpha(13\alpha-20)+8)\gamma}{\alpha-1} - 4\alpha + 4\right)$ which is positive for $c/V > L$, where both firms use autoscaling. Therefore, $F_{AA} - F_{NN}$ is positive

for $V > \frac{c}{\gamma}$, which includes all possible values of V in this case. Also for $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$, $F_{NA} - F_{NN}$ is increasing in V and equal to $\frac{c\gamma^2(\alpha(\alpha(\gamma-1)\gamma+2)-1)}{4(\alpha-1)((\gamma-1)\gamma+1)}$ at $V = \frac{c}{\gamma(1-\gamma)}$. Since $\alpha < \frac{1}{2}$, we have $\frac{c\gamma^2(\alpha(\alpha(\gamma-1)\gamma+2)-1)}{4(\alpha-1)((\gamma-1)\gamma+1)} > 0$, and therefore, $F_{NA} > F_{NN}$ for $V > \frac{c}{\gamma(1-\gamma)}$, which includes all possible values of V in this case.

Now consider when $\gamma > c/V$ and $\gamma(1-\gamma)V - c < 0$. In this case, $F_{AA} - F_{NN} = \gamma(V-c)(1 - \alpha - \gamma(1-2\alpha) - (\gamma V - c)/2$ which is convex in γ, equal to $-(V-c)(1-2\alpha)/2 < 0$ when evaluated at $\gamma = 1$, decreasing in γ through $\gamma = 1$, and equal to zero at

$$\hat{\gamma}_{AA} = \frac{V(1-2\alpha) - 2c(1-\alpha) + \sqrt{8c(V-c)(1-2\alpha) + (V - 2c(1-\alpha) - 2V\alpha)^2}}{4(V-c)(1-2\alpha)}.$$

Therefore, $F_{AA} - F_{NN}$ is negative for any $\gamma > \hat{\gamma}_{AA}$. Also in this case,

$$F_{NA} - F_{NN} = \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right) + \gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(\alpha-1)\gamma^2(c-V)} - \frac{1}{2}(\gamma V - c)$$

which is equal to $((-1+2\alpha)(c-V))/(4(-1+\alpha)) < 0$ at $\gamma = 1$ and $\frac{\partial(F_{NA}-F_{NN})}{\partial\gamma}$ equals $\frac{(2\alpha-1)(c-V)}{2(\alpha-1)} < 0$ at $\gamma = 1$. $F_{NA} = F_{NN}$ has one root between $\gamma = 0$ and $\gamma = 1$ which is

$$\hat{\gamma}_{NA} = 1/2\Big(- (((-1+\alpha)(1-\alpha+\sqrt{1-2\alpha})V)/(\alpha^2(c-V))) + \sqrt{\frac{(\alpha-1)\left(\alpha^3(V-2c)^2+\alpha^2\left(4\left(\sqrt{1-2\alpha}-1\right)c^2-4\left(\sqrt{1-2\alpha}-1\right)cV+\left(2\sqrt{1-2\alpha}-5\right)V^2\right)+2\left(3-2\sqrt{1-2\alpha}\right)\alpha V^2+2\left(\sqrt{1-2\alpha}-1\right)V^2\right)}{\alpha^4(c-V)^2}}\Big)$$

Therefore, $F_{NA} - F_{NN}$ is negative for any $\gamma > \hat{\gamma}_{NA}$. Thus, autoscaling decreases the range of F such that both firms enter the market for $\gamma > Max[\hat{\gamma}_{AA}, \hat{\gamma}_{NA}]$.

Proof of Corollary 1 in Essay 1

We show that for sufficiently small F (such that both firms enter the market), when $\gamma > c/V$, $\gamma(1-\gamma)V - c < 0$, and $\pi_{NA} < \pi_{AA} < \frac{\gamma V - c}{2}$, we have a prisoner's dilemma situation where both firms use autoscaling, even though their profits would be higher if autoscaling was not available.

When $\gamma > c/V$ and $\gamma(1-\gamma)V - c < 0$, both firms set their capacity to $\frac{1}{2}$, and each earn expected profit $\frac{\gamma V - c}{2}$, if autoscaling is not available. However, when autoscaling is available, since $\pi_{NA} < \pi_{AA}$, they both use autoscaling and each earn $\pi_{AA} < \frac{\gamma V - c}{2}$ in equilibrium, which creates the prisoner's dilemma. Now, we have to prove that all these conditions can be satisfied at the same time to show that the described region, in which prisoner's dilemma happens, actually exists.

Let $\gamma = \frac{1-\alpha}{2-3\alpha}$. After algebraic simplifications, we have both firms using autoscaling in equilibrium (i.e., $\pi_{NA} < \pi_{AA}$) if and only if $\frac{V}{c} > \frac{10\alpha^2-15\alpha+6}{4(\alpha-1)^2}$. Furthermore, using algebraic simplifications, the expected profit of autoscaling equilibrium for each firm is lower than that when firms do not use autoscaling (i.e., $\pi_{AA} < \frac{\gamma V-c}{2}$) if and only if $\frac{V}{c} < \frac{\alpha^2+2\alpha-2}{(\alpha-1)\alpha}$. It is easy to see that $\frac{10\alpha^2-15\alpha+6}{4(\alpha-1)^2} < \frac{\alpha^2+2\alpha-2}{(\alpha-1)\alpha}$ for any $\alpha \leq 1/2$. Therefore, for when $\frac{V}{c} \in (\frac{10\alpha^2-15\alpha+6}{4(\alpha-1)^2}, \frac{\alpha^2+2\alpha-2}{(\alpha-1)\alpha})$, the prisoner's dilemma situation holds.

Proof of Proposition 5 in Essay 1

If $c/V > L$ where L is defined in Lemma 2, then both firms would choose autoscaling upon entry and

$$E[CS_{AA}] = \begin{cases} \gamma^2(V-c)(1-2\alpha) & \text{if } F < (V-c)((1-\alpha)(1-\gamma)\gamma + \alpha\gamma^2) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

If $c/V < L$, then only one firm would choose autoscaling upon entry and

$$E[CS_{NA}] = \begin{cases} \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)\left(c\left((\alpha-2)\gamma^2-4(\alpha-1)\gamma+\alpha-1\right)+\gamma V(\alpha\gamma+\alpha-1)\right)}{4(\alpha-1)\gamma^2(c-V)} & \text{if } F < F_{NA} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $F_{NA} \equiv \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(\alpha-1)\gamma^2(c-V)}$.

Similarly, when autoscaling is not available, the expected consumer surplus is

$$E[CS_{NN}] = \begin{cases} \gamma^2 V(1-2\alpha) & \text{if } c < \gamma(1-\gamma)V \text{ and} \\ & V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))} \text{ and} \\ & F < (1-\alpha)(\gamma(1-\gamma)V-c)+\alpha\gamma^2 V \\ \frac{(1-\alpha)c^2}{4\gamma^2 V} + \frac{(\alpha-1)c(2\gamma+1)}{2\gamma} + \frac{V(\alpha(\gamma-1)+1)(\alpha\gamma+\alpha-1)}{4(\alpha-1)} & \text{if } c < \gamma(1-\gamma)V \text{ and} \\ & V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))} \text{ and} \\ & F < \frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Using Equations (8), (9), and (10), we compare consumer surplus with and without autoscaling. We start with the region $c < \gamma(1-\gamma)V$ and $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$, which represents Region 2 in Figure 5. Note that $V > \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$ and $0 < c < V$ require $\gamma < (1-\alpha)/(2-3\alpha)$, therefore based on Lemma (2) it is not possible that only one firm uses autoscaling in this region. Thus, we only compare cases when both firms use autoscaling with cases when autoscaling is not available. Using Equations (8) and (10), we can see that if $F < (1-\alpha)(\gamma(1-\gamma)V-c)+\alpha\gamma^2 V$, then both firms enter the market with or without autoscaling and $E[CS_{NN}] > E[CS_{AA}]$. For $(1-\alpha)(\gamma(1-\gamma)V-c)+\alpha\gamma^2 V < F < (V-c)((1-\alpha)(1-\gamma)\gamma+\alpha\gamma^2)$, $E[CS_{AA}] > E[CS_{NN}] = 0$. For $F > (V-c)((1-\alpha)(1-\gamma)\gamma+\alpha\gamma^2)$, $E[CS_{AA}] = E[CS_{NN}] = 0$.

Next, consider when $c < \gamma(1-\gamma)V$ and $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$, which represents Region 3 in Figure 5. First, we analyze the cases when both firms use autoscaling in Region 3. Comparing Equations (8) and (10) and assuming $c/V < \gamma$, we find that $E[CS_{NN}] > E[CS_{AA}]$ when $c/V < \frac{\gamma\left(4\alpha\gamma^3-\sqrt{\gamma}\sqrt{4(1-2\alpha)^2\gamma^5-8(\alpha-1)(2\alpha-1)\gamma^3-4(\alpha-1)(2\alpha-1)\gamma^2+(\alpha(13\alpha-20)+8)\gamma+4(\alpha-1)^2}-2\alpha\gamma-\alpha-2\gamma^3+2\gamma+1\right)}{1-\alpha}$ and $F < \frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)}$. Note that these conditions only hold when both firms use autoscaling, since $\frac{\gamma\left(4\alpha\gamma^3-\sqrt{\gamma}\sqrt{4(1-2\alpha)^2\gamma^5-8(\alpha-1)(2\alpha-1)\gamma^3-4(\alpha-1)(2\alpha-1)\gamma^2+(\alpha(13\alpha-20)+8)\gamma+4(\alpha-1)^2}-2\alpha\gamma-\alpha-2\gamma^3+2\gamma+1\right)}{1-\alpha}$ is positive if and only if $\gamma < (1-\alpha)/(2-3\alpha)$. Also in Region 3 of Figure 5, if both firms use autoscaling, $E[CS_{AA}] > E[CS_{NN}] = 0$ if $\frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)} < F < (V-c)((1-\alpha)(1-\gamma)\gamma+\alpha\gamma^2)$, and $E[CS_{AA}] = E[CS_{NN}] = 0$ if $F > (V-c)((1-\alpha)(1-\gamma)\gamma+\alpha\gamma^2)$.

Next, we compare the consumer surplus without autoscaling in Region 3, which occurs for

$c < \gamma(1-\gamma)V$ and $V < \frac{(1-\alpha)(2V\gamma^2+c)}{\gamma(1+\alpha(\gamma-1))}$, with consumer surplus when only one firm uses autoscaling, which occurs for $\gamma > (1-\alpha)/(2-3\alpha)$ and $c/V < L$. Comparing Equations (9) and (10), we find $E[CS_{NA}] > E[CS_{NN}]$ for $F < \frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)}$ and

$$
\begin{aligned}
c/V < & \frac{\alpha^2(2-\gamma(\gamma((\gamma-4)\gamma-4)+2))+2\alpha(\gamma(\gamma((\gamma-2)\gamma-5)+2)-2)+2(\gamma(3\gamma-1)+1)}{2(\alpha-1)^2} \\
& -((\alpha^2(2-\gamma(\gamma((\gamma-4)\gamma-4)+2))+2\alpha(\gamma(\gamma((\gamma-2)\gamma-5)+2)-2)+6\gamma^2-2\gamma+2)^2 \\
& +4(\alpha-1)^2\gamma(\alpha^2(\gamma^3-6\gamma^2+\gamma-2)+\alpha(4-2(\gamma-4)\gamma^2)-\gamma(2\gamma+1)-2))^{1/2}/(2(\alpha-1)^2)
\end{aligned}
$$

Since the term on the right hand side of the above inequality is greater than γ, thus for all $c/V < \gamma$, we have $E[CS_{NA}] > E[CS_{NN}]$. Furthermore in this region, $E[CS_{NA}] > E[CS_{NN}] = 0$ if $\frac{(\gamma V(1-\alpha(1-\gamma))-c(1-\alpha))^2}{4\gamma^2 V(1-\alpha)} < F < \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(\alpha-1)\gamma^2(c-V)}$, and $E[CS_{NA}] = E[CS_{NN}] = 0$ if $F > \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(\alpha-1)\gamma^2(c-V)}$.

Finally, if $c > \gamma(1-\gamma)V$, then when autoscaling is not available, both firms either do not enter the market or enter and set their prices equal to V. Both of these cases result in zero consumer surplus. Therefore autoscaling increases consumer surplus when both firms use autoscaling and $F < (V-c)((1-\alpha)(1-\gamma)\gamma+\alpha\gamma^2)$, or when one firm uses autoscaling and $F < \frac{\left(c\left(\alpha\left(\gamma^2-1\right)+1\right)+\gamma V(\alpha(-\gamma)+\alpha-1)\right)^2}{4(\alpha-1)\gamma^2(c-V)}$. Otherwise, autoscaling does not affect consumer surplus.

Proof that hosting all instances in the same pool when both users minimize is sub-optimal in Essay 2

Based on the decision timing shown in Figure 11, allocation decisions are made before interruptions are known. Here, we prove that the provider would optimally host each user on a different pool, when both users choose to minimize. Suppose the provider hosted all four instances in the same pool. The provider's expected profit from selling all four instances is $4(1-\gamma)\alpha v_l$, which is less than $E(\pi_{MM}) = 2(1-\gamma)\alpha(v_h+v_l)$. The profit from selling three instances depends on which of the values αv_h and v_l is bigger.

Suppose $\alpha v_h > v_l$. The expected profit from selling three instances is $3(1-\gamma)v_l$. For $\alpha v_h > v_l$, we have $3(1-\gamma)v_l < (2v_l+2\alpha v_l)(1-\gamma) < (2\alpha v_h+2\alpha v_l)(1-\gamma) = E(\pi_{MM})$. The expected profit from selling two instances is $2(1-\gamma)\alpha v_h$, which is also less than $E(\pi_{MM})$. Next, suppose

$\alpha v_h < v_l$. The expected profit from selling three instances is $3(1-\gamma)\alpha v_h$. For $\alpha v_h < v_l$, we have $3(1-\gamma)\alpha v_h < (v_l + 2\alpha v_h)(1-\gamma) < (2\alpha v_l + 2\alpha v_h)(1-\gamma) = E(\pi_{MM})$. The expected profit from selling two instances is $2(1-\gamma)v_l$, which is less than $2(1-\gamma)\alpha v_h$ for $\alpha v_h < v_l$, and is therefore less than $E(\pi_{MM})$. Therefore, the provider's expected profit is higher when each user is hosted on a different pool.

Proof of Proposition 1 in Essay 2

Throughout this proof we assume $\alpha > 2/3$. We compare the high value user's expected utilities $E(uHigh_{DM})$ with $E(uHigh_{MM})$, to find the optimal choice of allocation.

For $v_h < 3\alpha v_l$, we have $E(uHigh_{DM}) = (1-\gamma)^2\alpha(v_h - v_l) + \gamma(1-\gamma)(v_h - \alpha v_l)$. Thus for $v_h < 3\alpha v_l$, $E(uHigh_{DM}) > E(uHigh_{MM})$ if and only if $1 - \frac{\alpha(v_h - v_l)}{v_h(1-\alpha)} < \gamma < 1$.

For $3\alpha v_l < v_h < 3v_l$, we have $E(uHigh_{DM}) = (1-\gamma)^2\alpha(v_h - v_l)$. Thus for $3\alpha v_l < v_h < 3v_l$, $E(uHigh_{DM}) > E(uHigh_{MM})$ if and only if $0 < \gamma < 1 - \frac{v_h(1-\alpha)}{\alpha(v_h - v_l)}$. QED.

Proof of Proposition 2 in Essay 2

From proposition 1, we know for $v_h < 3\alpha v_l$ the high value user diversifies if and only if $1 - \frac{\alpha(v_h - v_l)}{v_h(1-\alpha)} < \gamma < 1$. The high value user's expected utility in this case is $E(uHigh_{DM}) = (1-\gamma)^2\alpha(v_h - v_l) + \gamma(1-\gamma)(v_h - \alpha v_l)$. If no interruption occurs, high value user's utility becomes $uHigh_{DM}^{AA} = \alpha(v_h - v_l)$. If the exclusive pool is interrupted, the price of the shared pool remains αv_l and the high value user's utility becomes $uHigh_{DM}^{IA} = v_h - \alpha v_l$. Thus, for $v_h < 3\alpha v_l$, we have $uHigh_{DM}^{IA} - uHigh_{DM}^{AA} = v_h(1-\alpha) > 0$. QED.

Proof of Proposition 3 in Essay 2

For $3\alpha v_l < v_h < 3v_l$, we have $E(uHigh_{DM}) = (1-\gamma)^2\alpha(v_h - v_l)$ and $E(uHigh_{MM}) = (1-\gamma)v_h(1-\alpha)$. Thus, $E(uHigh_{DM}) > E(uHigh_{MM})$ for $\alpha > \alpha^* = \frac{v_h}{v_h + (1-\gamma)(v_h - v_l)}$. For $\alpha > \alpha^*$, the high value user diversifies and the provider's expected profit is $E(\pi_{DM}) = (1-\gamma)^2(v_h + 3\alpha v_l) + 2\gamma(1-\gamma)v_h$. Since $E(\pi_{DM})$ is increasing in α, maximum profit for the provider with diversification is at $\alpha = 1$, which is $E(\pi_{DM}|\alpha \to 1) = (1-\gamma)^2(v_h + 3v_l) + 2\gamma(1-\gamma)v_h$. For $2/3 < \alpha < \alpha^*$, both users minimize and the provider's expected profit is $E(\pi_{MM}) = 2(1-\gamma)\alpha(v_h + v_l)$. The maximum profit for the

provider without diversification occurs at the highest α which is at the point of $\alpha = \alpha^*$, resulting in $E(\pi_{MM}|\alpha \to \alpha^*) = \frac{2v_h(1-\gamma)(v_h+v_l)}{v_h+(1-\gamma)(v_h-v_l)}$.

For $\gamma > \gamma^* = \frac{{v_h}^2-9v_hv_l+6{v_l}^2+\sqrt{v_h({v_h}^3-6{v_h}^2v_l+33v_h{v_l}^2-24{v_l}^3)}}{2{v_h}^2-8v_hv_l+6{v_l}^2}$, we have $E(\pi_{MM}|\alpha \to \alpha^*) > E(\pi_{DM}|\alpha \to 1)$, which means provider's expected profit is maximized at $\alpha = \alpha^*$. QED.