Efficient Capacity Provisioning for Firms with Multiple Locations: The Case of the Public Cloud

Patrick Hummel* and Michael Schwarz*

March 26, 2021

Abstract

This paper presents a model in which a firm with multiple locations strategically chooses capacity and prices in each location to maximize efficiency. We find that the firm provisions capacity in such a way that the probability an individual customer will be unable to purchase the goods the customer desires is lower in locations with greater expected demand. The firm also sets lower prices in larger locations. Finally, we illustrate that if a customer is indifferent between multiple locations, then it is more efficient to place this customer in a location with greater expected demand. These theoretical results are consistent with empirical evidence that we present from a major public cloud provider.

1 Introduction

There are a wide range of settings in which a firm has multiple locations of different sizes, and each of these different locations sells a homogeneous good. For example, many grocery store chains have multiple stores of different sizes that all sell the same groceries, and many restaurant chains have multiple different-sized restaurants that all have the same menus. In each of these settings, the chain must decide how much inventory to provide in each of its locations to meet the uncertain customer demand, what prices to charge, as well as how to advertise its locations to encourage customers to patronize one location or another.

The cloud computing market is another important example of such a setting. Major cloud providers such as Amazon Web Services, Microsoft Azure, and Google Cloud sell homogeneous cloud services in dozens of different regions throughout the world. In each of these regions, the cloud company provides computing capacity which can be rented on-demand for computation. Because the computing capacity can be rented on-demand, the cloud provider does not know what customer demand will be at any point in time, and the cloud provider must decide how much capacity to provision while taking into account the inherent uncertainty in customer demand. In addition, if a customer is indifferent between using multiple regions, the cloud company can encourage the customer to use whichever region would be most efficient.

*We thank Catherine de Fontenay, Catherine Wolfram, numerous colleagues at Microsoft, and seminar attendees at the Canadian Economic Theory Conference for helpful comments and discussions.
Corresponding Author: Patrick Hummel, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052. Email address: pahummel@microsoft.com

In each of these settings, the firm must provision capacity for its different locations while considering both the costs of provisioning capacity and the costs of not being able to meet customer demand if the uncertain demand exceeds capacity. These capacity choices will in turn impact expected utilization rates, the average fraction of capacity in a given location that the firm is able to sell, and this will be a key driver of cost differences between different locations. Utilization rates are important in many industries, but they are especially important in the cloud computing market, where capacity costs are a large fraction of overall costs.

In this type of setting, how should the firm provision capacity in different locations? How should the firm set prices in different locations? And if a firm can take actions that would steer customer demand towards one location or another, should the firm try to induce new demand to go to small locations or large locations?

Our interest in this problem was motivated by the following business question: should internal customers for Microsoft's cloud services be encouraged via internal pricing and other means to make use of regions with the lowest capacity utilizations? At first glance this may seem economically optimal. However, by analyzing the question more deeply, we discovered that the opposite strategy is optimal. Indeed, steering customers to regions with high utilization rates (which tend to be larger regions) can lead to noticeable cost savings even for a business of Microsoft's scale. While documenting the impact of steering internal Microsoft customers to larger regions is outside of the scope of this paper, this paper does provide theoretical and empirical results on efficient capacity provisioning and pricing when a firm has multiple locations.

We analyze a model in which a firm faces a competitive market and thus seeks to provision the efficient amount of capacity. In deciding how much capacity to provision in a region, the firm trades off the costs of providing additional capacity with the welfare gains associated with being able to satisfy additional demand if demand exceeds supply. Although we couch our model in terms of the cloud computing market, our results apply to any setting in which a firm has multiple locations of different sizes that sell a homogeneous good.

We show that when costs vary linearly with the amount of capacity provisioned, as the number of potential customers in a region increases, the firm provisions more capacity, and the expected fraction of demand that will be unfilled by the available capacity goes down. In addition, the price that is charged for compute also declines as a region becomes larger. This is consistent with empirical evidence that we present from Microsoft Azure.

Next, we address the question of whether it is more efficient to direct new customers who are willing to purchase compute in any region to a large region or a small region. In this setting, the expected fraction of capacity that will go unsold will be larger in small regions because the uncertainty in demand as a fraction of expected demand is larger in small regions. This implies that a supplier will overprovision capacity by a larger amount relative to expected demand in a small region than in a large region in order to maintain a high probability of being able to meet demand. Because a larger fraction of capacity is expected to be unsold in small regions, it seems logical to conjecture that it is more efficient to direct new demand to small regions in order to sell more of this unused capacity and help these regions achieve better economies of scale.

We show that this conjecture is false and that the opposite is true. The same reasoning that implies that a supplier will overprovision capacity by a larger amount relative to expected demand in small regions also implies that small regions have relatively higher average costs per unit of demand. Even though directing new demand to small regions

will help these regions achieve better economies of scale, the fact that small regions have relatively higher average costs per unit of demand implies that directing new demand to smaller regions will cause small regions to have to provision a larger amount of additional capacity as a result of the new demand than large regions. Thus, the marginal cost of serving an additional customer is larger in small regions than in large regions, and it is more efficient to direct new customers who are willing to purchase compute in any region to a larger region. This finding is also consistent with empirical evidence that we present from Microsoft Azure.

Finally, we extend our analysis to a model in which some customers are hyper-flexible in the sense that their workloads can be deployed in any region after observing the demand of other customers that must be placed in specific regions. This hyper-flexibility might be achieved in practice by offering a cloud product in which workloads can be migrated from one region to another after observing the arrival of new demand that must be placed in specific regions. Although such a product is not currently offered by the major cloud providers, cloud providers have expanded their technological capabilities over time, and it is possible that such a product might be developed in the future. It is thus worthwhile to understand how a cloud provider would provision capacity if customers can someday deploy these hyper-flexible workloads.

In this setting, we illustrate that when there is little hyper-flexible demand compared to the region-specific demand, the marginal cost of additional hyper-flexible demand will be many orders of magnitude smaller than the marginal cost of additional demand that must be placed in specific regions. Thus, introducing a hyper-flexible product would enable a cloud provider to significantly improve efficiency, even when the product is first introduced and there is relatively little adoption of the product.

However, when there is a large amount of hyper-flexible demand, there will be little difference between the marginal cost of additional hyper-flexible demand and the marginal cost of additional region-specific demand. Thus, if hyper-flexible cloud products eventually become sufficiently popular, it will make little difference to the cloud provider whether additional demand is hyper-flexible. However, the presence of a large amount of hyper-flexible demand will matter to customers in smaller regions because this hyper-flexible demand will lower the cost of compute in small regions closer to the price in large regions.

2 Background on Public Cloud

2.1 Industry Overview

The cloud computing industry is young, large, and rapidly growing. Although some of the concepts behind the public cloud were developed as early as the 1960s, all modern public clouds first emerged in the 21^{st} century (Foote 2017). Today annual world cloud revenues exceed $250 billion and are expected to grow by another 20% in 2021 (Graham *et al.* 2020a).

The public cloud consists of a wide range of services including infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). SaaS involves providing applications such as web-based email and productivity software to a consumer that can be accessed via the Internet. PaaS provides a platform for deploying consumer-created applications using the provider's programming languages, libraries, and tools. And IaaS provisions fundamental computing resources such as processing, storage, and

network to a consumer that can be used to deploy and run arbitrary software (Mell and Grance 2011).

Since the IaaS market can be thought of as a market for renting hardware, the questions on capacity provisioning studied in this paper are most relevant to the IaaS market, where annual world revenues currently exceed $50 billion (Graham *et al.* 2020a). The three largest IaaS providers in the United States are Amazon Web Services, Microsoft Azure, and Google Cloud, which account for 45%, 18%, and 5% of the world market respectively. In addition, the Chinese companies Alibaba and Tencent account for 9% and 3% of the world IaaS market respectively (Graham *et al.* 2020b).

Potential cloud customers can continue to use on-premise computing resources or can choose to move some of their computing needs to one of the public clouds. Many enterprise customers adopt a multi-cloud strategy of simultaneously using multiple different public cloud providers, simultaneously using both public cloud and private computing resources, or both (Flexera 2020). If a customer is initially using one public cloud, there are switching costs involved in moving to another public cloud, both due to the work involved in migrating particular applications and the fact that the customer will initially be less familiar and proficient with this other public cloud. Cloud providers attempt to reduce these switching costs by offering resources to help customers migrate from one public cloud to another (*e.g.* Microsoft Azure 2020c).

2.2 Virtual Machines

One of the main products sold by public cloud providers is a virtual machine or VM. From a user's point of view, a VM behaves like a computer that can be accessed remotely over the web. However, a VM is not a physically existing computer, as one physical machine can potentially power multiple small VMs (Microsoft Azure 2020f).

The largest cloud providers all sell a wide range of VMs that differ in dimensions such as the type of hardware that they use, whether they run Linux or Windows, their processing power, and the amount of memory (RAM) and temporary storage that they have. While there are differentiating factors between the major cloud providers, for most types of VMs offered by one of the largest public cloud providers, there are typically comparable offerings available from other large cloud providers at similar prices.[1]

2.3 Regions

Major cloud providers such as Amazon Web Services, Microsoft Azure, and Google Cloud each have dozens of regions around the world which house the physical machines which are used to provide cloud computing services to customers. For example, Microsoft Azure currently has over 60 announced regions, more than any other cloud provider (Microsoft Azure 2020e). In addition to many regions in the United States as well as many regions in various countries in Western Europe, Azure also has multiple regions in many other countries including Australia, Brazil, Canada, China, India, Japan, Korea, South Africa, United Arab Emirates, and others (Microsoft Azure 2020a).

In deploying a VM, a customer will be able to select the region for the deployment. There are several reasons customers may have preferences for particular regions. De-

[1]For example, the VMs offered by Amazon Web Services, Microsoft Azure, and Google Cloud can be found on https://aws.amazon.com/ec2/instance-types/, https://azure.microsoft.com/en-us/pricing/details/virtual-machines/series/, and https://cloud.google.com/compute/ respectively.

ploying to a more proximate region may reduce latency for a customer's applications. In addition, some customers may prefer to store data within a particular jurisdiction because of data custody laws. Finally, even if a customer is indifferent between multiple regions originally, after the customer has deployed a VM in a particular region, the customer may also prefer that any additional VMs are deployed in the same region because the customer is either storing data in that region or the customer has some critical process that runs continuously in that region.[2]

2.4 Business Models

Most cloud computing revenues come from pay as you go transactions, in which a customer only pays for the time in which the customer's VMs are deployed.[3] These list prices depend significantly on the type of VM that the customer uses and to a lesser extent on the region in which the VM is deployed. For example, on Microsoft Azure, the pay as you go price for a Linux A1v2 VM that runs on general purpose hardware with one virtual core, 2 gigabytes (GB) of RAM, and 10 GB of temporary storage ranges from $0.036 to $0.0793 per hour, depending on the region (Microsoft Azure 2020d). By contrast, the pay as you go price for a Linux M416msv2 VM that runs on specialized hardware with 416 virtual cores, 12 terabytes (TB) of RAM, and 8 TB of temporary storage is over $99 per hour in all regions (Microsoft Azure 2020d).[4]

While the list prices depend on the type of VM the customer uses as well as the region in which the VM is deployed, the list prices offered for such transactions tend to be stable over time. Although cloud providers do sometimes lower prices for VMs,[5] to the best of our knowledge, none of the major public clouds has ever increased prices for any of its VMs. In addition, these list prices do not change dynamically as a result of short-run changes in demand that may result during different times of day and different days of the week (Kilcioglu *et al.* 2017).

2.5 Why Auctions Are Not Used

The fact that compute is sold via fixed prices may seem surprising because demand varies over time as customers' computing needs evolve, and it may seem that an auction could more efficiently allocate capacity to customers when demand is uncertain and dynamically evolving. Why are auctions not used to allocate capacity to customers?

First, note that cloud providers provision enough capacity so that it is very unlikely that there will be a stockout in which there is not enough capacity to meet customer demand. Cloud customers may suffer substantial losses if a vital service is interrupted

[2]Wang *et al.* (2020) also empirically estimate a structural demand model of spatial competition using data from the cloud computing industry and find evidence that customers tend to prefer closer regions.

[3]Some customers also purchase compute via a reserved instance or savings plan, in which a customer commits to purchasing a given amount of compute in every hour for a period of 1-3 years, even if the customer did not deploy VMs during a particular hour. These reserved instances can result in discounts as large as 72% off the pay as you go price for three-year commitments on particular types of VMs (Microsoft Azure 2020b).

[4]It is worth noting that the difference in cost per operation for VMs with different specifications is typically smaller than the difference in price per hour because operations tend to finish more quickly on more powerful VMs. For example, Kilcioglu and Rao (2015) note that completing an operation with a VM whose price per hour is 8 times higher leads to a cost per operation that is 3 to 4 times higher.

[5]For example, in May 2017, Microsoft Azure announced price reductions of $4-7\%$ for general purpose virtual machines (Hillger 2017).

briefly because the customer was not able to purchase the VMs that it needs for a short period of time, much as an electricity consumer may suffer substantial losses if the consumer is unable to purchase needed electricity for a few hours.[6] Because cloud customers may suffer such substantial losses if they are unable to purchase the VMs that they need, cloud customers often have a value per unit of compute that is orders of magnitude higher than the corresponding capacity costs, and cloud providers try to ensure that with near certainty there will be enough capacity to meet demand.

Since cloud providers provision enough capacity to almost always be able to meet demand, if a cloud provider used an auction to sell compute to customers, the final price at the auction would almost always be equal to the reserve price. However, since cloud customers typically have a value per unit of compute that is orders of magnitude higher than the corresponding capacity costs, in the rare event that there was not enough capacity to meet all demand, the final price in an auction would be dramatically higher than the cloud provider's costs. Thus, if a cloud provider used an auction to sell compute to customers, there would be a very high probability that all customers could obtain all the compute they wanted at a low price and a low probability that the final price would be very high.

There are two problems with this pricing that would make auctions unsuitable in practice. First, using an auction results in a very high amount of uncertainty about the final realized prices. Thus, if either the cloud provider or the cloud customers are at all risk averse, using an auction to set prices will not meet either the cloud provider's or the cloud customers' needs.

Second, under an auction a cloud provider has a far stronger incentive to underinvest in capacity than under a fixed price mechanism. Under a fixed price mechanism, the cloud provider's revenue can only go down as a result of underinvesting in capacity, as the cloud provider will not be able to service as much demand. But under an auction, underinvesting in capacity will significantly increase a cloud provider's revenue by increasing the probability that there will not be enough capacity to meet demand, thereby increasing the probability that the final price in the auction will be very high. Thus, using a fixed price mechanism also enables a cloud provider to more credibly commit to provision the efficient amount of capacity. We illustrate these points formally in Appendix A in the paper.[7]

3 Related Literature

Our paper relates to several distinct strands of literature. First, there is a literature on pricing of cloud services (Abhisheki *et al.* 2012; Babaioff *et al.* 2017; Ben-Yehuda *et al.* 2013; Hoy *et al.* 2016; Kash and Key 2016; Kash *et al.* 2019; Kilcioglu *et al.* 2017). This literature largely focuses on questions related to comparing fixed and variable pricing for cloud services, but does not address questions related to pricing cloud services in different-sized regions, as we do in the present paper.

[6]There are some cloud customers who can tolerate interruptions of their workloads, and these customers may use Spot VMs, a VM that has a deep price discount, but can be evicted if the cloud provider needs to reclaim the capacity. Since Spot customers will not suffer substantial losses from having VMs evicted, the need for fixed prices does not apply to Spot VMs, and Spot VMs make use of time-varying prices (Shandilya 2020).

[7]In addition, Hummel (2018) illustrates in a formal model that a seller would have an incentive to provision less capacity under an auction than under fixed prices.

The operations research literature has also studied questions related to provisioning capacity for multiple locations. Much of this literature analyzes the economic benefits of risk pooling by consolidating multiple random demands into a single location. The earliest paper in this field is Eppen (1979), which illustrates in a newsboy model with normally distributed demands and linear costs that pooling multiple random demands leads to lower costs and that the cost difference is increasing in the variance of the demands but decreasing in the correlation between these demands. There are also a number of other papers that consider multi-location newsboy models such as Alfaro and Corbett (2003), Berman *et al.* (2011), Bimpkins and Markakis (2016), Chen and Lin (1989), Cherikh (2000), Gerchak and He (2003), Gerchak and Mossman (1992), and Yang and Schrage (2009) which extend Eppen's (1979) work in various ways. Finally, there has also been some work (*e.g.* Benjaafar *et al.* 2005) which analyzes the benefits of pooling in models of production-inventory systems.

Our work shares some features with this previous literature in that we also consider a newsboy model in which a supplier must provision capacity to meet an uncertain demand, and there is both a cost to provisioning capacity as well as a loss suffered from not having enough capacity to meet demand. The question we ask about how one should consolidate the demand from a new customer into the supplier's existing locations is also of the same flavor as the questions addressed by this risk pooling literature. However, the specific result we present about whether it is better to direct a new customer to a large location or a small location has not appeared in any of these previous papers.[8]

There has been comparably little theoretical work related to the results we present on how prices vary with the size of a firm's location. The only theory paper we are aware of that addresses the question of how prices vary with the size of a store is Braid (2003). This paper considers a model of spatial competition in which large stores alternate with small stores along an infinite roadway, and finds the opposite conclusion that larger stores will charge larger prices in equilibrium. Our model and results thus differ significantly from those in this previous paper.

Finally, there are some empirical papers that address questions related to capacity provisioning and pricing in different-sized grocery stores. Several empirical papers have found that the price of groceries tends to be lower at larger grocery stores (*e.g.* Alcaly and Klevorick 1971; Chung and Myers 1971; Kaufman 1998; Kaufman *et al.* 1997; Kunreuther 1972; Liese *et al.* 2007). These results give a specific empirical example of our theoretical finding that prices tend to be lower in larger locations. However, the mechanism driving these results could be different from the mechanism identified in our paper.

There is also evidence that larger grocery stores are less likely to run out of a particular type of grocery than smaller grocery stores, as Connell *et al.* (2007), Kaufman (1998), Kaufman *et al.* (1997), and Liese *et al.* (2007) have all found that larger grocery stores are more likely to have certain inventory than smaller grocery stores. These results are somewhat related to our theoretical finding that there is a lower probability that an individual customer will be unable to obtain the inventory the customer desires if the customer is in a larger location.

[8]Benjaafar *et al.* (2008) analyzes the problem of how to allocate demand that originates from multiple sources to different inventory locations, but again does not analyze whether it is better to direct a new customer to a large location or a small location.

4 Model

There are a total of N potential customers in a given region, each of whom demands some number of units of compute. Throughout we let D_i denote the demand of customer i. The demand of the customers, $(D_1, \ldots, D_N)$, is uncertain at the time the cloud provider provisions capacity to meet demand, but is known to be a random draw from some cumulative distribution function $G_N(D_1, \ldots, D_N)$.

If a customer wants a total of d units of compute, then the customer will be allocated no more than d units of compute. The customer then obtains a utility of kV if the customer is allocated a total of k units of compute, and a utility of 0 if the customer is not allocated any compute.

It costs the cloud provider a total of cQ to supply Q units of compute, where c is a cost parameter satisfying $c < V$. Because the cloud market is a competitive and rapidly growing market where a provider's long-run profits are likely to depend primarily on the amount of economic value created, we assume that the cloud provider chooses the capacity level Q to maximize efficiency.

For simplicity we also assume that the cloud provider sets a price p that results in zero expected profit. This last assumption is only used for one of the results in the paper, and this result will also hold under other plausible assumptions for how the cloud provider sets prices.

4.1 Assumptions on Demand Distribution

For arbitrary distributions of demand, $G_N(D_1, \ldots, D_N)$, it is difficult to make statements about how the price or the probability that an individual customer will fail to obtain a unit of compute that the customer desires will vary with N. Thus, we make some simplifying assumptions that are likely to hold in practice to assist with the analysis.

Throughout we assume that for sufficiently large values of N, the distribution of total demand, $D = \sum_{i=1}^{N} D_i$, is drawn from a continuous distribution $\Phi(D|\mu(N), \sigma(N))$ with mean $\mu(N)$ and standard deviation $\sigma(N)$, where $\Phi(D|\mu(N), \sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$ for some distribution $\Phi(\cdot)$ with mean 0 and standard deviation 1 that is symmetric about 0 in the sense that $\Phi(D) = 1 - \Phi(-D)$. We further assume that $\mu(N)$ and $\sigma(N)$ are increasing functions of N such that $\frac{\sigma(N)}{\mu(N)}$ is decreasing in N and $\sigma(N)$ is a strictly concave function of N.

This simplifying assumption will hold under many natural assumptions about customer demand. For example, if each customer's demand D_i is an independent and identically distributed draw from a distribution $G(\cdot)$ with bounded support, then for sufficiently large N, the distribution of customer demand is approximately normal with mean μN and standard deviation $\sigma\sqrt{N}$, where μ and σ denote the mean and standard deviation in the distribution $G(\cdot)$. Thus, this setting would satisfy the above assumption when $\Phi(\cdot)$ corresponds to a standard normal distribution, $\mu(N) = \mu N$, and $\sigma(N) = \sigma\sqrt{N}$.

In addition, this simplifying assumption also encompasses cases in which there can be systematic shocks to demand (*e.g.* a common component that impacts each of the customer demands $D_1, \ldots, D_N$), so $\frac{\sigma(N)}{\mu(N)}$ remains bounded away from zero, even in the limit as $N \to \infty$. Thus, this assumption is one that we could expect to hold in many practical settings.

5 Theoretical Results

5.1 How Price and Service Quality Vary with Region Size

This section addresses the question of how the prices and the probability that a customer will fail to obtain a unit of compute that the customer wants vary with N. We begin with the following preliminary lemma:

Lemma 1 *For sufficiently large values of N, the cloud provider sets a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$.*

All proofs are in Appendix B. The result in Lemma 1 implies that the cloud provider should provision an amount of capacity Q such that the probability there will not be enough capacity to meet demand is $\rho(Q) = \frac{c}{V}$ regardless of the size of the region N.

With this preliminary result in place, we now illustrate how the probability that an individual customer will fail to obtain a unit of compute that the customer wants varies with the size of the region N:

Theorem 1 *For sufficiently large values of N, the expected fraction of demand that will be unfilled by the available capacity is decreasing in N.*

Theorem 1 indicates that, even though the probability there will not be enough capacity to meet demand is the same in different-sized regions, the probability that an individual customer will fail to obtain a unit of compute that the customer wants will be lower in larger regions. The intuition for this result is that as N becomes larger, the amount of uncertainty in demand as a fraction of expected total demand declines. Thus, if demand exceeds supply, the expected difference between demand and supply as a fraction of total demand declines. This implies that the probability that an individual customer will fail to obtain a unit of compute that the customer wants will be lower in larger regions.

It is also worth noting that the probability an individual customer will fail to obtain a unit of compute that the customer wants may be much lower than the probability that there will not be enough capacity to meet demand. In order for a customer to fail to obtain a unit of compute that the customer wants, it is necessary for there to not be enough capacity to meet demand. But even if there is not enough capacity to fulfill all customer requests, it may be that there is enough capacity to fulfill the vast majority of customer requests. Thus, the probability an individual customer will fail to obtain a unit of compute that the customer wants may be much lower than the probability that there will not be enough capacity to meet demand.

We are also able to present results on how the price for compute varies with the size of the region:

Theorem 2 *For sufficiently large values of N, the price set for a unit of compute is decreasing in N.*

Since the cloud provider sets prices to achieve zero expected profit, the cloud provider will set prices in a region to reflect average costs. In larger regions, the amount of uncertainty as a fraction of expected total demand is lower, so the excess capacity needed (as a fraction of expected demand) to maintain a high probability of being able to meet all customer requests is also lower. Because of this, the expected fraction of capacity that

will go unused is smaller in larger regions, and average expected costs are also smaller in larger regions. Thus, the cloud provider can set lower prices in larger regions while still maintaining a non-negative profit margin. This explains the result in Theorem 2.

Theorem 2 was proven under the assumption that the cloud provider sets prices to achieve zero expected profit in each region, but analogs of this result will also hold under other plausible assumptions about how the cloud provider sets prices. As long as prices are chosen in such a way that prices will be correlated with average costs in a region, prices will tend to be lower in larger regions.

5.2 Selecting Regions for Customers

In this section we address the question of where a cloud provider should place customers that can be placed in any region. There are some customers that may have the flexibility to use any region, and when a cloud provider encounters such customers, the cloud provider must decide whether to encourage the customer to use a large region or a small region.

What is the most efficient way to direct demand from customers who can use any region? To answer this question, it is necessary to understand how adding demand to a region affects both the incremental capacity costs as well as the incremental number of deployment failures (*i.e.* the expected amount of demand that the cloud provider would fail to meet) in the region.

To address this question, we let $C(N)$ denote the capacity cost that is incurred in a region with N customers and $F(N)$ denote the expected number of deployment failures in a region with N customers. We then analyze how the incremental capacity cost, $C(N+1)-C(N)$, and the incremental expected number of deployment failures $F(N+1)-F(N)$, vary with the size of the region N. First we address this question for capacity costs:

Theorem 3 *Suppose the incremental increase in expected total demand when adding another customer to a region, $\mu(N+1)-\mu(N)$, is independent of N. Then the incremental capacity cost resulting from adding another customer to a region, $C(N+1)-C(N)$, is decreasing in N for sufficiently large values of N.*

The result in Theorem 3 implies that when there is a customer that has the flexibility to use any region, a cloud provider will incur a smaller incremental capacity cost if this customer is assigned to a larger region, as long as this customer would not change its expected demand as a result of being placed in the larger region. This result follows from the concavity of $\sigma(N)$. Because $\sigma(N)$ is concave in N, adding an additional customer to a larger region will do less to increase the amount of uncertainty in demand than adding this customer to a smaller region, and will thus also result in smaller incremental capacity costs in order to maintain the same probability of being able to meet all customer requests.

While adding an additional customer to a larger region results in smaller incremental capacity costs, it is worth noting that the percentage difference in incremental capacity costs between different-sized regions is likely to be small. Suppose, for example, that each customer's demand D_i is an independent and identically distributed draw from some cumulative distribution function $G(\cdot)$ with mean μ and standard deviation σ. In this case, we have $\mu(N) = \mu N$ and $\sigma(N) = \sigma\sqrt{N}$, so $\mu(N+1) - \mu(N) = \mu$ and $\sigma(N+1) - \sigma(N) = \frac{\sigma}{2\sqrt{N}} + O(\frac{1}{N})$ for sufficiently large N. Since we note in the proof of

Theorem 3 that $C(N+1)-C(N) = c[\mu(N+1)-\mu(N)+\Phi^{-1}(1-\frac{c}{V})(\sigma(N+1)-\sigma(N))]$, it then follows that $C(N+1)-C(N) = c[\mu+\Phi^{-1}(1-\frac{c}{V})(\frac{\sigma}{2\sqrt{N}}+O(\frac{1}{N}))]$.

For large values of N, this expression for $C(N+1)-C(N)$ will be within a few percent of $c\mu$, so the difference between the values of $C(N+1)-C(N)$ in two different-sized regions will be at most a few percent. Thus, the percentage difference in incremental capacity costs between different-sized regions would be small in this case.

Similarly, if there can be systematic shocks to demand, such as a common component that impacts each of the customer demands $D_1, \ldots, D_N$, in addition to these idiosyncratic demand differences between different customers, then we might have $\sigma(N) = \alpha N + \sigma\sqrt{N}$ for some positive constants α and σ in addition to $\mu(N) = \mu N$. In this case, we would have $C(N+1) - C(N) = c[\mu(N+1) - \mu(N) + \Phi^{-1}(1-\frac{c}{V})(\sigma(N+1)-\sigma(N))] = c[\mu+\Phi^{-1}(1-\frac{c}{V})(\alpha+\frac{\sigma}{2\sqrt{N}}+O(\frac{1}{N}))]$. Similar reasoning would then imply that the percentage difference in incremental capacity costs between two different-sized regions is likely to be no more than a few percent.

Next we address the question of how the size of the region where we place excess demand impacts the expected number of deployment failures:

Theorem 4 *Let $F(N)$ denote the expected number of deployment failures that are incurred in a region with N customers. Then the incremental expected number of deployment failures resulting from adding another customer to a region, $F(N+1) - F(N)$, is decreasing in N for sufficiently large values of N.*

The result in Theorem 4 further implies that when there is a customer that has the flexibility to use any region, a cloud provider will incur fewer incremental deployment failures if this customer is assigned to a larger region. Furthermore, unlike the case of incremental capacity costs considered in Theorem 3, the percentage difference in the incremental expected number of deployment failures resulting from adding a customer to a different-sized region may be substantial.

In the proof of Theorem 4, we note that $F(N) = \int_{\Phi^{-1}(1-\frac{c}{V})}^{\infty}(z-\Phi^{-1}(1-\frac{c}{V}))\sigma(N)\, d\Phi(z)$, so $F(N+1)-F(N)$ is proportional to $\sigma(N+1)-\sigma(N)$. Thus, the ratio between the incremental expected number of deployment failures resulting from adding another customer to a region with N customers and the incremental expected number of deployment failures resulting from adding another customer to a region with $2N$ customers is $\frac{\sigma(N+1)-\sigma(N)}{\sigma(2N+1)-\sigma(2N)}$.

Now we have seen previously that if each customer's demand D_i is an independent and identically distributed draw from some cumulative distribution function $G(\cdot)$ with standard deviation σ, then $\sigma(N+1)-\sigma(N) = \frac{\sigma}{2\sqrt{N}}+O(\frac{1}{N})$ for sufficiently large N. Thus in this particular case, the ratio $\frac{\sigma(N+1)-\sigma(N)}{\sigma(2N+1)-\sigma(2N)} \approx \frac{2\sigma\sqrt{2N}}{2\sigma\sqrt{N}} = \sqrt{2}$, which implies that adding another customer to a region with N customers instead of $2N$ customers results in $\sqrt{2}$ times as many incremental deployment failures, or roughly 40% more incremental deployment failures. Thus, adding a new customer to a larger region can result in significantly better quality of service than adding this customer to a smaller region.

Finally, we can combine the results in Theorems 3 and 4 to illustrate that it is more efficient to place new customers who can use any region in larger regions:

Theorem 5 *If a new customer can be placed in any region, it is most efficient to place this customer in the largest region possible. This will result in the lowest incremental capacity costs and the smallest number of incremental deployment failures.*

5.3 Differences Between Marginal Costs and Average Costs

In deciding whether to direct new customers to a large region or a small region, a firm needs to weigh the fact that small regions have larger average costs per unit demand along with the fact that the additional demand is more likely to help a small region achieve economies of scale. The results in the previous section indicate that this trade-off always resolves in such a way that it is more efficient to direct new demand to large regions.

But although Theorem 3 indicates that marginal costs are higher in small regions than in large regions, one might wonder if the fact that additional demand is more likely to help a small region achieve economies of scale implies that the difference in marginal costs between small regions and large regions is smaller than the difference in average costs. Is there a theoretical guarantee of this?

It appears that the answer to this question is no. However, there are some important special cases of the model in which we do have a guarantee that the difference in marginal cost between different-sized regions is smaller than the difference in average costs. This section presents analysis that compares the differences in marginal costs with the differences in average costs for different-sized regions.

Suppose there are two regions, A and B, and region A has N^* customers while region B has βN^* customers for some $\beta > 1$. To answer the question of how the difference in marginal costs between regions A and B compares to the difference in average costs, we first provide mathematical expressions for these differences:

Lemma 2 *Suppose that expected total demand for a region with N customers, $\mu(N) = \mu N$ for some constant $\mu > 0$. Then the difference in average costs between regions A and B is $c\Phi^{-1}(1 - \frac{c}{V})(\frac{\sigma(N^*)}{N^*} - \frac{\sigma(\beta N^*)}{\beta N^*})$, while the difference in marginal costs between these regions is $c\Phi^{-1}(1 - \frac{c}{V})[(\sigma(N^* + 1) - \sigma(N^*)) - (\sigma(\beta N^* + 1) - \sigma(\beta N^*))]$.*

With these expressions for the differences in average costs and marginal costs in mind, we first note why there is no general guarantee that the difference in marginal costs between regions A and B will be smaller than the difference in average costs:

Example 1 *Suppose that $\sigma(N) = \gamma N$ for $N \leq \beta N^*$ and $\sigma(N)$ is independent of N for values of $N \geq \beta N^*$. Then the difference in marginal costs between regions A and B is greater than the corresponding difference in average costs.*

The particular example presented in Example 1 could arise when there is both uncertainty and perfect correlation between the demands of the first βN^* customers, but there is no uncertainty about the demands of any additional customers. This example illustrates that it is at least theoretically possible for the difference in marginal costs between different-sized regions to be greater than the difference in average costs.[9]

But while it is theoretically possible for the difference in marginal costs between different-sized regions to be greater than the difference in average costs, this will not arise in some other situations that are more likely to occur in practice. We illustrate that the difference in marginal costs between different-sized regions is guaranteed to be less than the difference in average costs in an important special case of the model:

[9]Strictly speaking, this example only satisfies the weaker conditions that $\frac{\sigma(N)}{\mu(N)}$ is non-increasing in N and $\sigma(N)$ is concave in N rather than the stronger conditions that $\frac{\sigma(N)}{\mu(N)}$ is decreasing in N and $\sigma(N)$ is strictly concave in N stated in Section 4.1. However, if we slightly perturb $\sigma(N)$ when $N \leq \beta N^*$ to make $\sigma(N)$ a strictly concave function, $\sigma(N)$ will satisfy the conditions in Section 4.1, and the conclusion in Example 1 will still hold.

Theorem 6 *Suppose that each customer's demand D_i is an independent and identically distributed draw from some cumulative distribution function $G(\cdot)$ with mean μ and standard deviation σ. Then the difference in marginal costs between regions A and B is half as large as the corresponding difference in average costs.*

Theorem 6 indicates that in an important special case of the model, directing a new customer to a small region instead of a large region does indeed do more to help the small region achieve economies of scale. Thus, the differences in marginal costs between small regions and large regions are smaller than the differences in average costs in this setting.

6 Hyper-Flexible Customers

We now extend the model to a situation in which some customers are hyper-flexible in the sense that their workloads can be deployed in any region after observing the demand of other customers that must be placed in specific regions. In order to analyze the consequences of this hyper-flexible demand, we modify the model to explicitly consider multiple different regions as well as both region-specific and hyper-flexible demand.

6.1 Extended Model

We suppose there are R regions, and we let N_r denote the number of customers who must be assigned to region r. We also let $D_{i,r}$ denote the demand from customer i in region r and let $D_r \equiv \sum_{i=1}^{N_r} D_{i,r}$ denote the total demand from customers who must be assigned to region r. Finally, for each region r, we assume that total demand in region r, D_r, is drawn from a distribution $\Phi(D_r|\mu(N_r), \sigma(N_r))$ with mean $\mu(N_r)$ and standard deviation $\sigma(N_r)$, where $\Phi(\cdot)$, $\mu(N)$, and $\sigma(N)$ satisfy the same properties given in Section 4.1.

To model hyper-flexible customers, we let N_{flex} denote the number of customers who are hyper-flexible, and let $D_{i,flex}$ denote the demand from hyper-flexible customers. We also let $D_{flex} \equiv \sum_{i=1}^{N_{flex}} D_{i,flex}$ denote the total demand from hyper-flexible customers, and we assume throughout that D_{flex} is known.

The customer's utility function remains the same as the original model. Thus, if a customer wants a total of d units of compute, the customer will be allocated no more than d units of compute, the customer will obtain a utility of kV if the customer obtains a total of k units of compute, and a utility of 0 if the customer is not allocated any compute. This holds regardless of whether the customer has hyper-flexible demand.

Likewise, as in the original model, the cloud provider's costs vary linearly with the amount of capacity supplied. Thus, if the cloud provider supplies Q_r units of capacity in region r, then the cloud provider's total capacity costs are $c\sum_{r=1}^{R} Q_r$ for some cost parameter $c < V$. The cloud provider again chooses a level of capacity Q_r in each region r to maximize efficiency.

6.2 Results

Let $\rho_r(Q_1, \ldots, Q_R)$ denote the probability that either (i) there will not be enough capacity in region r to meet the region-specific demand in region r or (ii) there will not be enough supply left over to meet the hyper-flexible demand after placing as much of the region-specific demand in the corresponding regions as possible. That is, $\rho_r(Q_1, \ldots, Q_R)$ denotes the probability that either $D_r > Q_r$ or $D_{flex} > \sum_{r=1}^{R} \max\{Q_r - D_r, 0\}$.

In this case, the marginal value of an additional unit of capacity in region r is $\rho_r(Q_1, \ldots, Q_R)V$, so in order to maximize efficiency, the cloud provider needs to choose the values of $Q_1, \ldots, Q_R$ to satisfy $\rho_r(Q_1, \ldots, Q_R)V = c$. Thus, we have the following lemma:

Lemma 3 *There exist efficiency-maximizing levels of capacity $Q_1, \ldots, Q_R$ in regions $1, \ldots, R$. These values are chosen to satisfy $\rho_r(Q_1, \ldots, Q_R) = \frac{c}{V}$.*

We now turn to the question of how the amount of capacity that the cloud provider supplies to meet additional hyper-flexible demand compares to the amount of capacity that must be provisioned to meet region-specific demand. The amount of additional capacity needed to meet additional region-specific demand will surely be at least as large as the amount of additional capacity needed to meet additional hyper-flexible demand, but how big will the difference be?

Since the answer to this question may depend significantly on the context, we begin by studying a situation in which the amount of hyper-flexible demand is small relative to the amount of demand that must be place in specific regions and the demand in different regions is independently distributed. In this setting, we prove the following result:

Theorem 7 *Suppose that demand in each region r, D_r, is an independently distributed random variable. If there is no hyper-flexible demand, then for sufficiently large values of N_r, the marginal cost of servicing additional hyper-flexible demand is $R(\frac{c}{V})^{R-1}$ times the marginal cost of servicing additional demand that must be placed in a specific region.*

The result in Theorem 7 implies that if there is little hyper-flexible demand, then the marginal cost of servicing additional hyper-flexible demand is many orders of magnitude smaller than the marginal cost of servicing additional region-specific demand. It is very rare for major cloud providers not to have enough capacity to meet demand in a particular region, so we should expect the value of $\frac{c}{V}$, the probability that there will not be enough capacity to meet demand in a particular region, to be very small in practical applications. Given this, the value of $R(\frac{c}{V})^{R-1}$ in the statement of Theorem 7 will be very low, and the marginal cost of servicing additional hyper-flexible demand will be many orders of magnitude smaller than the marginal cost of servicing additional region-specific demand.

The intuition for this result is as follows: If there is additional demand that must be placed in a specific region, then the cloud provider must supply additional capacity in that region to fulfill that demand. But if there are multiple regions, then in the process of supplying capacity for the region-specific demand, it is very likely that the cloud provider will supply more capacity than is needed for the region-specific demand in some region. This excess capacity can in turn be used to fulfill the hyper-flexible demand. Thus, with probability close to 1, the cloud provider will be able to satisfy the hyper-flexible demand, even without supplying additional capacity. For this reason, the cost of servicing a small amount of hyper-flexible demand is very small compared to the cost of servicing a similar amount of demand that must be placed in a specific region.

But while the cost of servicing additional hyper-flexible demand is close to zero in the setting in Theorem 7, this result hinges crucially on the assumptions that the demand in different regions is independently distributed and that there is little hyper-flexible demand compared to the region-specific demand. The difference in marginal cost between hyper-flexible and region-specific VMs is driven by the fact that it is far less likely that we will not have excess capacity in some region that can be used to fulfill additional hyper-flexible

demand than it is that we will not have excess capacity in a specific region that can be used to fulfill additional region-specific demand. However, this difference becomes smaller without the assumptions of independently distributed demand and little hyper-flexible demand compared to the region-specific demand.

First, when there is stronger correlation in demand between the different regions, the difference between the probability that we will not have excess capacity in some region that can be used to fulfill additional hyper-flexible demand and the probability that we will not have excess capacity in a specific region that can be used to fulfill additional region-specific demand becomes smaller. When demand in the different regions is more correlated, there is a higher chance that we will not have excess capacity in other regions when we do not have excess capacity in a particular region, so the differences in these probabilities becomes smaller. Thus, the difference in marginal cost between hyper-flexible and region-specific VMs also becomes smaller when demand in different regions becomes more correlated. Indeed, in the extreme case where demand in the different regions is perfectly correlated, the difference in marginal cost between hyper-flexible and region-specific VMs would vanish.

In addition, the difference between the probability that we will not have excess capacity in some region that can be used to fulfill additional hyper-flexible demand and the probability that we will not have excess capacity in a specific region that can be used to fulfill additional region-specific demand also becomes smaller when there is more hyper-flexible demand. When the amount of hyper-flexible demand becomes large relative to the amount of region-specific demand, the marginal cost for additional hyper-flexible demand becomes the same as the marginal cost for additional demand that must be placed in a specific region. We prove this result below in a setting where we assume that upper bound of the support of D_r, $\overline{D_r}$, is finite for all regions r:[10]

Theorem 8 *For sufficiently large amounts of hyper-flexible demand, the marginal cost of servicing additional hyper-flexible demand equals the marginal cost of servicing additional demand that must be placed in a specific region.*

To understand the intuition for Theorem 8, note that if there is a large amount of hyper-flexible demand, then the cloud provider will have to supply significantly more capacity than what is needed solely to meet the region-specific demand. For this reason, it is virtually certain that there will be enough capacity in the individual regions to meet the region-specific demand, and the cloud provider will only need to focus on supplying enough total capacity to meet total demand. But if the cloud provider is simply choosing total capacity levels to meet total demand, then there will be no difference between the marginal cost of additional region-specific demand and the marginal cost of additional hyper-flexible demand. This explains the result in Theorem 8.

7 Empirical Results

This section presents empirical results that measure the extent to which our theoretical results identified in Section 5 hold empirically. We use data from Microsoft Azure to

[10] Although an analogous result holds without this assumption, making this assumption enables us to greatly simplify the exposition of the proof. In the context of the public cloud, each customer has some finite limit on the maximum number of cores that the customer can deploy at any given time. Thus, when applying the model to the public cloud, the upper bound of the support of D_r will be finite in practice.

illustrate the extent to which price and incremental capacity costs vary with the size of the region.

7.1 Prices

First, we analyze how prices vary with the size of a region in Azure. Throughout this section we use data from the wide variety of VMs that a customer can purchase. Even within a given region, a customer has the flexibility to deploy different types of VMs that meet a customer's needs. For example, Azure currently offers virtual machines that are general purpose (such as Dv3), compute optimized (such as Fsv2), and memory optimized (such as Ev3), as well as many others (Microsoft Azure 2020e).

Because Azure offers such a wide range of different types of VMs, not all VMs can run on the same hardware. This means that the total amount of supply that is available for one type of VM in a region may differ from the total amount of supply that is available for another type of VM in a region. In addition, the total demand for one type of VM in a region may differ from the total demand for another type of VM in the same region.

Due to the above considerations, in defining the size of a region, we use definitions that capture the fact that a region may be bigger for one type of VM than for another. In particular, we define the total supply for a particular type of VM in a region as the total number of physical cores in the region that could be used to host this type of VM. We also define the total demand for a particular type of VM in a region as the total number of physical cores that customers demanded for that type of VM at a point in time.

When addressing the question of how the price for a unit of compute varies with the size of the region, we then use a definition of region size that is particular to the type of VM in question. Throughout we find that the supply-based and demand-based measures of region size are nearly perfectly correlated, so we just report the results using the supply-based measure of region size. The results for the demand-based measure of region size are nearly identical.

For each of six different types of Azure VMs, we noted the price, total supply, and total demand for this VM type in each region that was available to a US customer as of March 2020 (44 regions in total). We then analyzed the degree of correlation between the price and total supply that could be used to host this type of VM across the various regions. We also depict a scatterplot of VM prices and region size for these six VM types in Figure 1.

The results of this analysis reveal significant negative correlation between the price and the size of the region. For each of the types of VMs in question, we estimate a correlation coefficient between price and region size that falls somewhere between -0.38 and -0.48, with an average correlation coefficient of -0.43. We estimate a correlation coefficient between price and the log of the region size that falls somewhere between -0.37 and -0.57, with an average correlation coefficient of -0.48. And we estimate a correlation coefficient between the log of price and the log of the region size that falls somewhere between -0.37 and -0.60, with an average correlation coefficient of -0.50.

Finally, the average prices for these types of VMs were consistently $10-20\%$ higher in the smallest $\frac{1}{3}$ of regions than in the largest $\frac{1}{3}$ of regions. These results thus reveal that there is significant negative correlation between price and region size in practice, consistent with the theoretical predictions in Theorem 2.

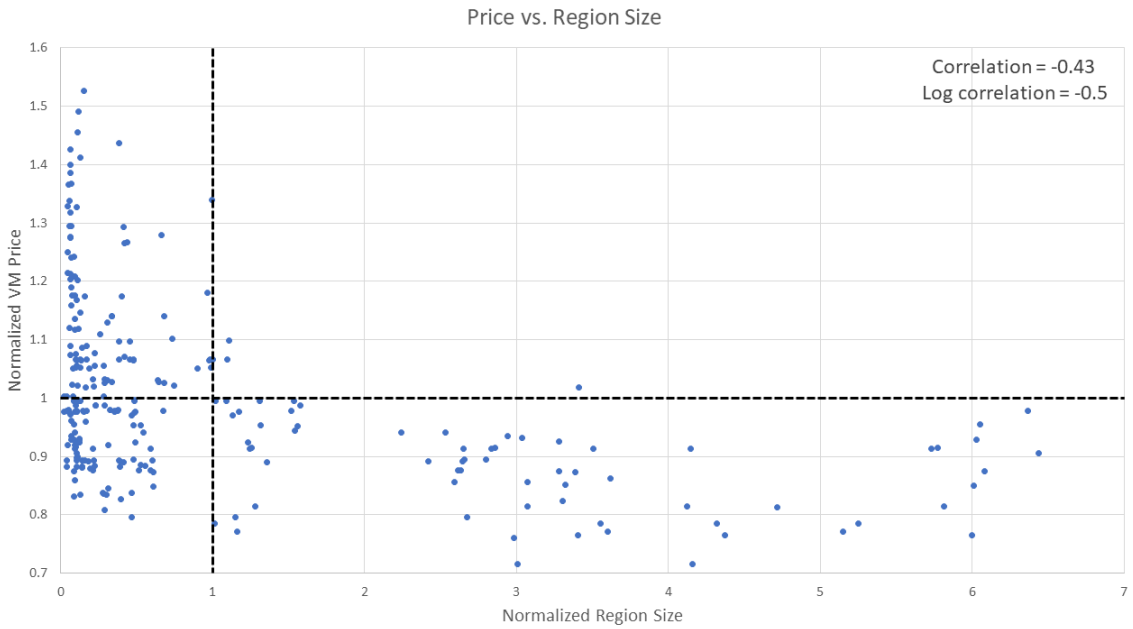

Figure 1: Scatterplot of the normalized VM price compared to the normalized supply for that VM type in different regions. This scatterplot is based on six different types of VMs in Microsoft Azure, and the normalized price (supply) for a particular VM type in a particular region is defined to be the ratio of the price (supply) for that VM type in that region to the average price (supply) for that VM type over all regions. Of the 70 cases where a region had more than the average supply for a particular VM type (cases to the right of the vertical dashed line), only 4 were cases where this region had a higher price than average (cases above the horizontal dashed line), and the average price for these 70 cases was 11% lower than the overall average. By contrast, of the 194 cases where a region had less than the average supply for a particular VM type, these regions had a higher price than average half the time, and sometimes the price was 50% higher than the overall average. The correlation between normalized VM price and normalized region size is -0.43 and the correlation between the log of these quantities is -0.5.

7.2 Incremental Capacity Costs

Next, we analyze how the amount of capacity that is supplied per additional unit of demand varies with the size of a region. This section provides an empirical illustration of our theoretical result in Theorem 3 that the incremental capacity cost from adding more demand to a region is decreasing in the size of the region.

In analyzing this problem, we account for the fact that a given type of capacity in Azure can be used by multiple different types of VMs, so capacity must be supplied to meet total demand for these different types of VMs, rather than merely separately ensuring that there is enough capacity for each individual type of VM. We thus make use of a formulation of capacity supplied and demand that reflects these subtleties.

In particular, for each country where some region was available to a US customer from December 2019 to November 2020, as well as each day t in this time frame, we calculate the total supply in that country that could be used to host some subset of general purpose VMs in Azure, Q_t. We also calculate the total demand in that country for the types of VMs that could make use of some of this particular type of capacity, D_t.

After collecting these data points (D_t, Q_t) for every day t in a given country, we then run a linear regression of Q_t on D_t. The coefficient in this regression β_c for a particular country c then gives a measure of the average amount of additional capacity that is supplied for each additional unit of demand in that country.

After calculating the coefficient in this regression separately for each country, we then measured the correlation between the coefficient in this regression and the average size of the regions within the country. As in Section 7.1, the supply-based and demand-based measures of region size are nearly perfectly correlated, so we just report the results using the supply-based measure of region size. We also calculate 95% confidence intervals in these correlation coefficients by generating $10,000$ bootstrap samples of countries and calculating correlation coefficients for each of these bootstrap samples.

The results of this analysis revealed significant negative correlation between the coefficients estimated in this regression and the size of the regions within the country. We estimate a correlation coefficient of -0.30 between β_c and region size (with a 95% confidence interval of $[-0.53, -0.11]$), -0.36 between β_c and the log of region size (with a 95% confidence interval of $[-0.69, 0.03]$), and -0.37 between $\log(\beta_c)$ and the log of region size (with a 95% confidence interval of $[-0.72, 0.06]$). Thus, there is negative correlation between the incremental capacity cost from adding more demand to a region and the size of the region, consistent with the theoretical predictions in Theorem 3.

8 Conclusion

Although there are many practical settings in which a firm with multiple locations must strategically provision capacity and set prices in different-sized locations, there has been little work that addresses the question of the most efficient way for a firm to achieve these objectives. This paper has analyzed this question and shown that a firm should provision capacity in such a way that it is less likely that an individual customer will be unable to purchase the goods the customer desires in a region with greater expected demand. The firm should also set lower prices in its locations with greater capacity and expected demand. Finally, the firm should steer customers who are willing to purchase from multiple locations to its larger locations.

While the results in this paper can be applied to many settings in which a firm provisions capacity for multiple locations, they are especially relevant for the cloud computing market, where major cloud providers typically supply cloud services in dozens of different regions throughout the world. Our theoretical findings on how marginal costs and prices vary with the size of a region are consistent with practice at Microsoft Azure, as marginal costs and prices tend to be higher in Azure's smaller regions.

Appendix

Appendix A: Consequences of Using Auctions

This appendix analyzes the consequences of using an auction to sell compute to customers. We consider a special case of the model presented in Section 4 in which each customer either demands 0 or 1 units of compute. As noted in the discussion of Lemma 1, the cloud provider chooses the level of capacity Q so that the probability demand exceeds supply is $\frac{c}{V}$.

If the cloud provider uses a second-price auction with a reserve price of r, then it is a dominant strategy for each customer to make a bid equal to the customer's value for a unit of compute, and all customers who want a unit of compute will bid V for a unit of compute. With probability $(1 - \frac{c}{V})$, demand will be less than supply and any capacity sold will be sold at a per unit price of r. But with probability $\frac{c}{V}$, demand will be greater than supply and all capacity will be sold at a per unit price of V, the common bid of all the bidders in the auction.

For the public cloud, the probability that there will not be enough capacity to meet demand, $\frac{c}{V}$, will typically be less than 0.01. Thus, if there is not enough capacity to meet demand, the final price at the auction V will be over 100 times larger than the cost per unit of capacity c. By contrast, if there is enough capacity to meet demand, the final price at the auction will be equal to the reserve price r, which we can expect to be the same order of magnitude as c.

These results indicate that if a cloud provider uses an auction to sell compute to customers, then there will be tremendous uncertainty about the final price in the auction. Thus, if either the cloud provider or the cloud customers are risk averse, using an auction to sell compute to customers would not meet either party's needs.

In addition, the cloud provider has much stronger incentives to underinvest in capacity under an auction than under a fixed price mechanism. Suppose, for example, the cloud provider reduces the amount of capacity provisioned to a level so that the probability there is not enough supply to meet demand is $\frac{2c}{V}$ rather than $\frac{c}{V}$. Under a fixed price mechanism, this will reduce the cloud provider's expected revenue because it will reduce the maximum amount of demand the cloud provider can service, while having no impact on prices.

But under an auction, the expected price in the auction will increase from $(1 - \frac{c}{V})r + \frac{c}{V}V = (1-\frac{c}{V})r+c$ to $(1-\frac{2c}{V})r+\frac{2c}{V}V = (1-\frac{2c}{V})r+2c$ as a result of this decrease in capacity. For reserve prices r ranging from 0 to c, this could range from a $50-100\%$ increase in the expected price. Thus, underinvesting in capacity could significantly increase expected prices in an auction, and thereby significantly increase the cloud provider's revenue. This in turn implies that the cloud provider would have significantly greater incentives to underinvest in capacity under an auction than under a fixed price mechanism.

Appendix B: Proofs of Main Results

Proof of Lemma 1. Let $\rho(Q)$ denote the probability that there will not be enough capacity to meet demand for all customers for a given capacity choice Q. In this case, the marginal value of an additional unit of capacity to customers is $\rho(Q)V$, so in order to maximize efficiency, the cloud provider needs to choose Q in such a way that $\rho(Q)V = c$, meaning we would have $\rho(Q) = \frac{c}{V}$.

Since the distribution of total demand, $D = \sum_{i=1}^{N} D_i$, is drawn from the distribution $\Phi(D|\mu(N), \sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$ for sufficiently large values of N, in order to ensure that the probability there will not be enough capacity to meet demand is $\rho(Q)$, the cloud provider should set $Q = \mu(N) + \Phi^{-1}(1 - \rho(Q))\sigma(N)$. This implies that in order to maximize efficiency, the cloud provider should choose $Q = \mu(N) + \Phi^{-1}(1-\rho(Q))\sigma(N) = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$. □

Proof of Theorem 1. For sufficiently large N, the distribution of total demand, $D = \sum_{i=1}^{N} D_i$, is drawn from the distribution $\Phi(D|\mu(N), \sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$. Furthermore, we know from Lemma 1 that the cloud provider sets $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$.

If the cloud provider sets this level of capacity, then the expected fraction of demand that will be unfilled by the available capacity is $\int_{\Phi^{-1}(1-\frac{c}{V})}^{\infty} \frac{(z-\Phi^{-1}(1-\frac{c}{V}))\sigma(N)}{\mu(N)+z\sigma(N)}\, d\Phi(z) = \int_{\Phi^{-1}(1-\frac{c}{V})}^{\infty} \frac{z-\Phi^{-1}(1-\frac{c}{V})}{(\mu(N)/\sigma(N))+z}\, d\Phi(z)$. Since $\frac{\sigma(N)}{\mu(N)}$ is decreasing in N, it follows that $\frac{\mu(N)}{\sigma(N)}$ is increasing in N and $\int_{\Phi^{-1}(1-\frac{c}{V})}^{\infty} \frac{z-\Phi^{-1}(1-\frac{c}{V})}{(\mu(N)/\sigma(N))+z}\, d\Phi(z)$ is decreasing in N. Thus, the expected fraction of demand that will be unfilled by the available capacity is decreasing in N. □

Proof of Theorem 2. Since the cloud provider sets a price that will result in zero expected profit, the cloud provider sets a price p so that $pE[\min\{D, Q\}] = cQ$, where $D = \sum_{i=1}^{N} D_i$ denotes the uncertain realization of total demand and Q denotes the cloud provider's capacity choice. Thus, the price for a unit of compute is decreasing in N if and only if $\frac{Q}{E[\min\{D,Q\}]}$ is decreasing in N, which is equivalent to $E[\min\{\frac{D}{Q}, 1\}]$ being increasing in N. We thus seek to prove that $E[\min\{\frac{D}{Q}, 1\}]$ is increasing in N.

We know that for sufficiently large N, the distribution of total demand is drawn from the distribution $\Phi(D|\mu(N), \sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$. In addition, we know from Lemma 1 that the cloud provider sets $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$. Thus, under these circumstances, $E[\min\{\frac{D}{Q}, 1\}] = \int_{-\infty}^{\infty} \frac{\mu(N)+\min\{z, \Phi^{-1}(1-\frac{c}{V})\}\sigma(N)}{\mu(N)+\Phi^{-1}(1-\frac{c}{V})\sigma(N)}\, d\Phi(z)$.

Since $\Phi(\cdot)$ is symmetric about 0, we know that $\int_{-\infty}^{\infty} \min\{z, \Phi^{-1}(1 - \frac{c}{V})\}\, d\Phi(z) = -K$ for some constant $K > 0$ that is independent of N. Thus, $E[\min\{\frac{D}{Q}, 1\}] = \frac{\mu(N)-K\sigma(N)}{\mu(N)+\Phi^{-1}(1-\frac{c}{V})\sigma(N)} = \frac{\mu(N)/\sigma(N)-K}{\mu(N)/\sigma(N)+\Phi^{-1}(1-\frac{c}{V})}$ for some constant $K > 0$. Since $\frac{\sigma(N)}{\mu(N)}$ is decreasing in N, it follows that $\frac{\mu(N)}{\sigma(N)}$ is increasing in N and $E[\min\{\frac{D}{Q}, 1\}] = \frac{\mu(N)/\sigma(N)-K}{\mu(N)/\sigma(N)+\Phi^{-1}(1-\frac{c}{V})}$ is increasing in N. Thus, the price for a unit of compute is decreasing in N. □

Proof of Theorem 3. We know from Lemma 1 that for sufficiently large N, the cloud provider sets $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$. Thus, $C(N) = c[\mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)]$ gives the capacity cost that is incurred in a region with N customers.

The above result in turn implies that $C(N+1) - C(N) = c[\mu(N+1) - \mu(N) + \Phi^{-1}(1 - \frac{c}{V})(\sigma(N+1) - \sigma(N))]$. Since $\mu(N+1) - \mu(N)$ is independent of N, it then follows that $C(N+1) - C(N)$ is decreasing in N if and only if $\sigma(N+1) - \sigma(N)$ is decreasing in

N. And since $\sigma(N)$ is a strictly concave function of N, we know that $\sigma(N+1)-\sigma(N)$ is indeed decreasing in N. Thus, the incremental capacity cost resulting from adding another customer to a region, $C(N+1)-C(N)$, is decreasing in N. $\square$

Proof of Theorem 4. We know that for sufficiently large N, the distribution of total demand, $D=\sum_{i=1}^{N} D_i$, is drawn from the distribution $\Phi(D|\mu(N),\sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$. In addition, we know from Lemma 1 that the cloud provider sets $Q=\mu(N)+\Phi^{-1}(1-\frac{c}{V})\sigma(N)$. Thus, under these circumstances, the expected number of deployment failures would be $F(N)=\int_{\Phi^{-1}(1-\frac{c}{V})}^{\infty}(z-\Phi^{-1}(1-\frac{c}{V}))\sigma(N)\,d\Phi(z)$.

The above result in turn implies that $F(N+1)-F(N)=\int_{\Phi^{-1}(1-\frac{c}{V})}^{\infty}(z-\Phi^{-1}(1-\frac{c}{V}))(\sigma(N+1)-\sigma(N))\,d\Phi(z)$. Since $\sigma(N)$ is a strictly concave function of N, it follows that $\sigma(N+1)-\sigma(N)$ is decreasing in N, and thus that this expression for $F(N+1)-F(N)$ is decreasing in N. $\square$

Proof of Theorem 5. It follows from Theorem 3 that placing a new customer in the largest region will result in the lowest incremental capacity costs and from Theorem 4 that placing a new customer in the largest region will result in the smallest number of incremental deployment failures. Thus, if a new customer can be placed in any region, it is most efficient to place this customer in the largest region. $\square$

Proof of Lemma 2. We know from the proof of Theorem 3 that the capacity cost for a region with N customers is $C(N)=c[\mu(N)+\Phi^{-1}(1-\frac{c}{V})\sigma(N)]$, so when $\mu(N)=\mu N$ for some constant $\mu>0$, we have $\frac{C(N)}{N}=c[\mu+\Phi^{-1}(1-\frac{c}{V})\frac{\sigma(N)}{N}]$. From this it follows that the difference in average costs between regions A and B is $\frac{C(N^*)}{N^*}-\frac{C(\beta N^*)}{\beta N^*}=c\Phi^{-1}(1-\frac{c}{V})(\frac{\sigma(N^*)}{N^*}-\frac{\sigma(\beta N^*)}{\beta N^*})$.

Similarly, we know from Theorem 3 that if $\mu(N)=\mu N$ for some constant $\mu>0$, then the marginal cost of serving a region with N customers is $C(N+1)-C(N)=c[\mu+\Phi^{-1}(1-\frac{c}{V})(\sigma(N+1)-\sigma(N))]$. Thus, the difference in marginal costs between regions A and B is $[C(N^*+1)-C(N^*)]-[C(\beta N^*+1)-C(\beta N^*)]=c\Phi^{-1}(1-\frac{c}{V})[(\sigma(N^*+1)-\sigma(N^*))-(\sigma(\beta N^*+1)-\sigma(\beta N^*))]$. $\square$

Proof of Example 1. If $\sigma(N)=\gamma N$ for $N\le\beta N^*$ and $\sigma(N)$ is independent of N for values of $N\ge\beta N^*$, then $\frac{\sigma(N)}{N}-\frac{\sigma(\beta N)}{\beta N}=\gamma-\gamma=0$, while $(\sigma(N+1)-\sigma(N))-(\sigma(\beta N+1)-\sigma(\beta N))=\gamma-0=\gamma$. Since $(\sigma(N+1)-\sigma(N))-(\sigma(\beta N+1)-\sigma(\beta N))=\gamma$ is greater than $\frac{\sigma(N)}{N}-\frac{\sigma(\beta N)}{\beta N}=0$, it follows that the difference in marginal costs between regions A and B is greater than the corresponding difference in average costs. $\square$

Proof of Theorem 6. If each customer's demand D_i is an independent and identically distributed draw from some cumulative distribution function $G(\cdot)$ with mean μ and standard deviation σ, then $\mu(N)=\mu N$ and $\sigma(N)=\sigma\sqrt{N}$. Thus, $\frac{\sigma(N)}{N}-\frac{\sigma(\beta N)}{\beta N}=\frac{\sigma}{\sqrt{N}}-\frac{\sigma}{\sqrt{\beta N}}=\frac{\sigma(\sqrt{\beta}-1)}{\sqrt{\beta N}}$, while $(\sigma(N+1)-\sigma(N))-(\sigma(\beta N+1)-\sigma(\beta N))=\frac{\sigma}{2\sqrt{N}}-\frac{\sigma}{2\sqrt{\beta N}}+O(\frac{1}{N})=\frac{\sigma(\sqrt{\beta}-1)}{2\sqrt{\beta N}}+O(\frac{1}{N})$. Since $(\sigma(N+1)-\sigma(N))-(\sigma(\beta N+1)-\sigma(\beta N))=\frac{\sigma(\sqrt{\beta}-1)}{2\sqrt{\beta N}}$ is half as large as $\frac{\sigma}{\sqrt{N}}-\frac{\sigma}{\sqrt{\beta N}}=\frac{\sigma(\sqrt{\beta}-1)}{\sqrt{\beta N}}$, it follows that the difference in marginal costs between regions A and B is half as large as the corresponding difference in average costs. $\square$

Proof of Lemma 3. We have already seen that any efficiency-maximizing levels of capacity $Q_1,\ldots,Q_R$ must satisfy $\rho_r(Q_1,\ldots,Q_R)=\frac{c}{V}$, so it is sufficient to prove that

efficiency-maximizing levels of capacity exist. To see this, let Q^* denote some value of Q such that the probability that total demand $D_{flex} + \sum_{r=1}^{R} D_r$ exceeds Q^* is less than $\frac{c}{V}$. For any values of $Q_r \geq Q^*$, the marginal value of an additional unit of capacity in region r will always be less than the marginal cost of this capacity, so the cloud provider will never want to supply more than Q^* units of capacity in region r. Thus, the cloud provider will seek to find values of $Q_1, \ldots, Q_R$ that maximize efficiency subject to the constraint that $Q_r \leq Q^*$ for all r.

But the problem of maximizing efficiency subject to the constraint that $Q_r \leq Q^*$ for all r involves maximizing a continuous function of $(Q_1, \ldots, Q_R)$ over a closed and bounded set. Since there exists a solution to any such optimization problem, it follows that there exist efficiency-maximizing levels of capacity $Q_1, \ldots, Q_R$ in regions $1, \ldots, R$. $\square$

Proof of Theorem 7. If there is no hyper-flexible demand, then we know from Lemma 1 that the cloud provider chooses Q_r in each region r so that $Q_r = \mu(N_r) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N_r)$. This in turn implies that the cost incurred in region r is $C_r = c[\mu(N_r) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N_r)]$, meaning the marginal cost of servicing ϵ more demand that must be placed in region r is $c\epsilon$, as the value of the expression for C_r increases by $c\epsilon$ when $\mu(N_r)$ increases by ϵ.

Now suppose we must service ϵ more hyper-flexible demand for some arbitrarily small $\epsilon > 0$. In the limit as $\epsilon \to 0$, the probability that we will have enough excess capacity in a particular region r to meet this hyper-flexible demand after allocating the region-specific demand is $\frac{c}{V}$. And since there are a total of $R - 1$ other regions besides region r, the probability that we will have enough excess capacity in some region other than region r to meet this hyper-flexible demand after allocating the region-specific demand is $(\frac{c}{V})^{R-1}$.

The above results imply that if there is ϵ more hyper-flexible demand for some arbitrarily small $\epsilon > 0$, then with probability $(\frac{c}{V})^{R-1}$ the cloud provider will need to allocate ϵ more demand to region r, and with probability $1 - (\frac{c}{V})^{R-1}$ the cloud provider will not need to allocate any additional demand to region r. Thus, having ϵ more hyper-flexible demand increases expected demand in region r, $\mu(N_r)$, by $(\frac{c}{V})^{R-1}\epsilon$.

In addition, since the variance in the amount of hyper-flexible demand that must be allocated to region r is $O(\epsilon^2)$, having ϵ more hyper-flexible demand increases the standard deviation of the demand that must be placed in region r, $\sigma(N_r)$ by $O(\epsilon^2)$. And since the cost of servicing demand in region r is $C_r = c[\mu(N_r) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N_r)]$, having ϵ more hyper-flexible demand increases the value of the expression for C_r by $c(\frac{c}{V})^{R-1}\epsilon + O(\epsilon^2)$, meaning the marginal cost of servicing ϵ more hyper-flexible demand in region r is $c(\frac{c}{V})^{R-1}\epsilon + O(\epsilon^2)$.

Since there are R regions, this means that the total marginal cost of servicing ϵ more hyper-flexible demand in all regions is $cR(\frac{c}{V})^{R-1}\epsilon + O(\epsilon^2)$, which is $R(\frac{c}{V})^{R-1}$ times as large as $c\epsilon$ in the limit as $\epsilon \to 0$. Since the marginal cost of servicing ϵ more demand that must be placed in region r is $c\epsilon$, this in turn implies that the marginal cost of servicing additional hyper-flexible demand is $R(\frac{c}{V})^{R-1}$ times the marginal cost of servicing additional demand that must be placed in a specific region. $\square$

Proof of Theorem 8. We have seen in the proof of Theorem 7 that the marginal cost of servicing an additional ϵ demand that must be placed in a specific region is $c\epsilon$. Thus, it suffices to prove that, if the amount of hyper-flexible demand D_{flex} is sufficiently large, the marginal cost of servicing an additional ϵ hyper-flexible demand is $c\epsilon$. To prove this, we first prove that if $Q_1, \ldots, Q_R$ denote the optimal capacity choices in a given region,

then for sufficiently large values of D_{flex}, the probability that we will not have enough capacity to meet region-specific demand, $Pr(D_r > Q_r)$, is 0 in all regions r.

To see this, recall from Lemma 3 that the values of $Q_1, \ldots, Q_R$ must be chosen in such a way that $\rho_r(Q_1, \ldots, Q_R) = \frac{c}{V}$. Since $\rho_r(Q_1, \ldots, Q_R)$ is greater than or equal to the probability that $D_{flex} > \sum_{r=1}^{R} \max\{Q_r - D_r, 0\}$, this in turn implies that these values must be chosen in such a way that $Pr(D_{flex} > \sum_{r=1}^{R} \max\{Q_r - D_r, 0\}) \leq \frac{c}{V}$. This further implies that the values of $Q_1, \ldots, Q_R$ must be chosen in such a way that $Pr(D_{flex} > \sum_{r=1}^{R} Q_r - D_r) \leq \frac{c}{V}$, which in turn holds if and only if $Pr(D_{flex} + \sum_{r=1}^{R} D_r > \sum_{r=1}^{R} Q_r) \leq \frac{c}{V}$.

Since it is necessary for $Pr(D_{flex} + \sum_{r=1}^{R} D_r > \sum_{r=1}^{R} Q_r) \leq \frac{c}{V}$ to hold under an optimal capacity provisioning strategy, for values of $D_{flex} > \sum_{r=1}^{R} \overline{D_r}$, the value of $\sum_{r=1}^{R} Q_r$ must exceed $\sum_{r=1}^{R} \overline{D_r}$ as well. Thus, if we let Q^* denote the optimal total amount of capacity to purchase (*i.e.* $Q^* \equiv \sum_{r=1}^{R} Q_r$), then for values of $D_{flex} > \sum_{r=1}^{R} \overline{D_r}$, it will be possible to choose values of $Q_1, \ldots, Q_R$ in such a way that $\sum_{r=1}^{R} Q_r = Q^*$ and $Pr(D_r > Q_r) = 0$ for all regions r.

Now suppose the cloud provider chooses the amounts of capacity to purchase in each region, $Q_1, \ldots, Q_R$, in such a way that $\sum_{r=1}^{R} Q_r = Q^*$. If the cloud provider also chooses these values of $Q_1, \ldots, Q_R$ to satisfy $Pr(D_r > Q_r) = 0$ for all regions r, then the cloud provider will maximize the total amount of demand that the cloud provider will be able to service for the following reason:

The total amount of region-specific demand that the cloud provider can service is $D_{spec} = \sum_{r=1}^{R} \min\{D_r, Q_r\}$ and the total amount of hyper-flexible demand that the cloud provider can service is $\min\{D_{flex}, Q^* - D_{spec}\}$. Thus, the total amount of demand that the cloud provider can service is $D_{spec} + \min\{D_{flex}, Q^* - D_{spec}\} = \min\{D_{spec} + D_{flex}, Q^*\}$, which is increasing in D_{spec}.

But $D_{spec} = \sum_{r=1}^{R} \min\{D_r, Q_r\} \leq \sum_{r=1}^{R} D_r$ for any values of $Q_1, \ldots, Q_R$. And if $Q_1, \ldots, Q_R$ are chosen in such a way that $Pr(D_r > Q_r) = 0$ for all regions r, then $D_{spec} = \sum_{r=1}^{R} D_r$ regardless of the realizations of $D_1, \ldots, D_R$. Thus, D_{spec} is maximized by choosing the values of $Q_1, \ldots, Q_R$ in such a way that $Pr(D_r > Q_r) = 0$ for all regions r, so for any fixed value of Q^*, the cloud provider will maximize the total amount of demand that can be serviced by choosing the values of $Q_1, \ldots, Q_R$ in such a way that $Pr(D_r > Q_r) = 0$ for all regions r.

Now recall from Lemma 3 that the cloud provider chooses the values of $Q_1, \ldots, Q_R$ in such a way that $\rho_r(Q_1, \ldots, Q_R) = \frac{c}{V}$ for all regions r. Since $\rho_r(Q_1, \ldots, Q_R)$ denotes the probability that either $D_r > Q_r$ or $D_{flex} > \sum_{r=1}^{R} \max\{Q_r - D_r, 0\}$, and we know that the cloud provider chooses $Q_1, \ldots, Q_R$ in such a way that $Pr(D_r > Q_r) = 0$ for all regions r, it follows that $\rho_r(Q_1, \ldots, Q_R) = Pr(D_{flex} > \sum_{r=1}^{R} Q_r - D_r)$ for all r. This in turn implies that the cloud provider chooses the values of $Q_1, \ldots, Q_R$ in such a way that $Pr(D_{flex} > \sum_{r=1}^{R} Q_r - D_r) = Pr(D_{flex} + \sum_{r=1}^{R} D_r > \sum_{r=1}^{R} Q_r) = \frac{c}{V}$.

But if the value of D_{flex} increases by ϵ, then the value of $\sum_{r=1}^{R} Q_r$ that is needed to ensure that $Pr(D_{flex} + \sum_{r=1}^{R} D_r > \sum_{r=1}^{R} Q_r) = \frac{c}{V}$ also increases by ϵ. Thus, if the amount of hyper-flexible demand D_{flex} is sufficiently large, the marginal cost of servicing an additional ϵ hyper-flexible demand is $c\epsilon$ and thus equal to the marginal cost of servicing an additional ϵ demand that must be placed in a specific region. $\square$

References

[1] Abhishek, Vineet, Ian A. Kash, and Peter Key. 2012. "Fixed and Market Pricing for Cloud Services". *2012 Proceedings IEEE INFOCOM Workshops*. 157–162.

[2] Alcaly, Roger E. and Alvin K. Klevorick. 1971. "Food Prices in Relation to Income Levels in New York City". *Journal of Business*. 44(4): 380-397.

[3] Alfaro, José A. and Charles J. Corbett. 2003. "The Value of SKU Rationalization in Practice (The Pooling Effect Under Suboptimal Inventory Policies and Nonnormal Demand)". *Production and Operations Management*. 12(1): 12-29.

[4] Babaioff, Moshe, Yishay Mansour, Noam Nisan, Gali Noti, Carlo Curino, Nar Ganapathy, Ishai Menache, Omer Reingold, Moshe Tennenholtz, and Erez Tinmat. 2017. "ERA: A Framework for Economic Resource Allocation for Cloud". *Proceedings of the* 26^{th} *International Conference on World Wide Web*. 635-642.

[5] Benjaafar, Saif, William L. Cooper, and Joon-Seok Kim. 2005. "On the Benefits of Pooling in Production-Inventory Systems". *Management Science*. 51(4): 548-565.

[6] Benjaafar, Saif, Yanzhi Li, Dongsheng Xu, and Samir Elhedhli. 2008. "Demand Allocations in Systems with Multiple Inventory Locations and Multiple Demand Sources". *Manufacturing & Service Operations Management*. 10(1): 43-60.

[7] Ben-Yehuda, Orna Agmon, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. 2013. "Deconstructing Amazon EC2 Spot Instance Pricing". *ACM Transactions on Economics and Computation*. Article No. 16.

[8] Berman, Oded, Dmitry Krass, and M. Mahdi Tajbakhsh. 2011. "On the Benefits of Risk Pooling in Inventory Management". *Production and Operations Management*. 20(1): 57-71.

[9] Bimpikis, Kostas and Mihalis G. Markakis. 2016. "Inventory Pooling Under Heavy-Tailed Demand". *Management Science*. 62(6): 1533-1841.

[10] Braid, Ralph M. 2003. "Spatial Price Competition Between Large and Small Stores with Stockouts or Limited Product Selections". *Economics Letters*. 81(2): 257-262.

[11] Chen, Maio-Sheng and Chin-Tsai Lin. 1989. "Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem". *Journal of the Operational Research Society*. 40(6): 597-602.

[12] Cherikh, M. 2000. "On the Effect of Centralisation on Expected Profits in a Multi-Location Newsboy Problem". *Journal of the Operational Research Society*. 51: 755-761.

[13] Chung, Chanjin and Samuel L. Myers Jr. 1971. "Do the Poor Pay More for Food? An Analysis of Grocery Store Availability and Food Price Disparities". *Journal of Consumer Affairs*. 33(2): 276-296.

[14] Connell, Carol L., M. Kathleen Yadrick, Pippa Simpson, Jeffrey Gossett, Bernestine McGee, and Margaret L. Bogle. 2007. "Food Supply Adequacy in the Lower Mississippi Delta". *Journal of Nutritional Education and Behavior*. 39(2): 77-83.

[15] Dierks, Ludwig and Sven Seuken. 2019. "Cloud Pricing: The Spot Market Strikes Back". *Proceedings of the 2019 Conference on Economics and Computation*. 593-594.

[16] Eppen, Gary D. 1979. "Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem". *Management Science*. 25(5): 498-501.

[17] Flexera. 2020. "Flexera 2020 State of the Cloud Report".

[18] Foote, Keith D. 2017. "A Brief History of Cloud Computing". https://www.dataversity.net/brief-history-cloud-computing/.

[19] Gerchak, Yigal and Qi-Ming He. 2003. "On the Relation Between the Benefits of Risk Pooling and the Variability of Demand". *IIE Transactions*. 35(11): 1027-1031.

[20] Gerchak, Yigal and David Mossman. 1992. "On the Effect of Demand Randomness on Inventories and Costs". *Operations Research*. 40(4): 633-825.

[21] Graham, Colleen, Fabrizio Biscotti, Bindi Bhullar, Vanitha Dsilva, Neha Gupta, and Terilyn Palanca. 2020a. "Forecast: Public Cloud Services, Worldwide, 2018-2024, 2Q20 Update". Gartner Technical Report ID G00717080.

[22] Graham, Colleen, Fabrizio Biscotti, Bindi Bhullar, Terilyn Palanca, Craig Roth, Neha Gupta, Julian Poulter, Alys Woodward, and Jim Hare. 2020b. "Market Share: Enterprise Public Cloud Services, Worldwide, 2019". Gartner Technical Report ID G00717082.

[23] Hillger, Brian. 2017. "Price Reductions on General Purpose Virtual Machines". Microsoft Azure Blog (May 19, 2017).

[24] Hoy, Darrell, Nicole Immorlica, and Brendan Lucier. 2016. "On-Demand or Spot? Selling the Cloud to Risk-Averse Customers". *Proceedings of the* 12^{th} *International Conference on Web and Internet Economics*. 73-86.

[25] Hummel, Patrick. 2018. "How Do Selling Mechanisms Affect Profits, Surplus, Capacity, and Prices with Unknown Demand?". *Canadian Journal of Economics* 51(1): 94-126.

[26] Kash, Ian A. and Peter B. Key. 2016. "Pricing the Cloud". *IEEE Internet Computing* 20(1): 36–43.

[27] Kash, Ian A., Peter B. Key, and Warut Suksompong. 2019. "Simple Pricing Schemes for the Cloud". *ACM Transactions on Economics and Computation*. Article No. 7.

[28] Kaufman, Phil R. 1998. "Rural Poor Have Less Access to Supermarkets, Large Grocery Stores". *Rural Development Perspectives*. 13(3): 19-26.

[29] Kaufman, Phil R., James M. MacDonald, Steve M. Lutz, and David M. Smallwood. 1997. "Do the Poor Pay More for Food? Item Selection and Price Differences Affect Low-Income Household Food Costs". US Department of Agriculture Agricultural Economic Report No. 759.

[30] Kilcioglu, Cinar and Justin M. Rao. 2015. "Competition on Price and Quality in Cloud Computing". *Proceedings of the* 25^{th} *International Conference on the World Wide Web* 1123-1132.

[31] Kilcioglu, Cinar, Justin M. Rao, Aadharsh Kannan, and R. Preston McAfee. 2017. "Usage Patterns and the Economics of the Public Cloud". *Proceedings of the* 26^{th} *International Conference on the World Wide Web*. 83-91.

[32] Kunreuther, Howard. 1973. "Why the Poor May Pay More for Food: Theoretical and Empirical Evidence". *Journal of Business*. 46(3): 368-383.

[33] Liese, Angela D., Kristina E. Weis, Delores Pluto, Emily Smith, and Andrew Lawson. 2007. "Food Store Types, Availability, and Cost of Foods in a Rural Environment". *Journal of the American Dietic Association*. 107(11): 1916-16923.

[34] Mell, Peter and Timothy Grance. 2011. "The NIST Definition of Cloud Computing". National Institute of Standards of Technology Special Publication 800-145.

[35] Microsoft Azure. 2020a. "Azure Geographies". https://azure.microsoft.com/en-us/global-infrastructure/geographies/.

[36] Microsoft Azure. 2020b. "Azure Reserved VM Instances (RIs)". https://azure.microsoft.com/en-us/pricing/reserved-vm-instances/

[37] Microsoft Azure. 2020c. "Discover, Assess, and Migrate Amazon Web Services (AWS) VMs to Azure". https://docs.microsoft.com/en-us/azure/migrate/tutorial-migrate-aws-virtual-machines

[38] Microsoft Azure. 2020d. "Linux Virtual Machines Pricing". https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/

[39] Microsoft Azure. 2020e. "Virtual Machine Series". https://azure.microsoft.com/en-us/pricing/details/virtual-machines/series/.

[40] Microsoft Azure. 2020f. "What is a Virtual Machine?". https://azure.microsoft.com/en-us/overview/what-is-a-virtual-machine/.

[41] Shandilya, Varun. 2020. "Announcing the General Availability of Spot Virtual Machines". Microsoft Azure Blog (May 12, 2020).

[42] Wang, Shuang, Jacob LaRiviere, and Aadharsh Kannan. 2020. "Spatial Competition and Missing Data: An Application to Cloud Computing." Boston University Working

Paper.

[43] Yang, Hongsuk and Linus Schrage. 2009. "Conditions that Cause Risk Pooling to Increase Inventory". *European Journal of Operational Research*. 192(3): 837-851.