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

Patrick Hummel* and Michael Schwarz*

July 20, 2020

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 it is more efficient to direct customers who are willing to go to multiple locations to locations with greater expected demand.

1 Introduction

There are a wide range of settings in which a firm has multiple locations of different sizes that sell a homogeneous good, and the firm must decide how much capacity to provision given uncertain demand, what prices to charge, and whether to encourage customers who are willing to go to multiple locations to go to one of the firm's larger or smaller locations. 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 capacity to provision in each of its locations to meet the uncertain customer demand as well as how to advertise its locations to encourage customers to attend one location or another.

Another important example of such a setting is the cloud computing market. 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.

*Microsoft Corporation, One Microsoft Way, Redmond, WA 98052. Email addresses: Patrick Hummel: pahummel@microsoft.com; Michael Schwarz: mschwarz@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. How should the firm provision capacity in the 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?

This paper analyzes these questions in the context of a model in which a firm faces a competitive market and thus seeks to provision an amount of capacity to maximize efficiency while setting prices that result in zero expected profit. 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 illustrate that when costs vary linearly with the amount of capacity provisioned, as the number of potential customers in a region becomes larger, the firm provisions capacity in such a way that (i) the probability the firm provisions enough capacity for all customers stays the same and (ii) the probability any individual customer will fail to obtain a unit of compute when the customer wants it goes down. In addition, the price that is charged for compute also declines as a region becomes larger, consistent with empirical evidence from Microsoft Azure.

Finally, 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. Here a firm faces a trade-off in deciding whether to direct new demand to small regions or large regions. Small regions have larger average costs per unit demand because the uncertainty in demand as a fraction of expected demand is larger in small regions, so it is necessary to provision more capacity per unit of expected demand in order to maintain a high probability of being able to meet demand in a small region. At the same time, a small region will also benefit more from additional demand because the additional demand will do more to help this region achieve economies of scale. We show that this trade-off always resolves in such a way that it is more efficient to direct new demand to large regions.

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.

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 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 in this general area such as Benjaafar *et al.* (2008), 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) that extend Eppen's (1979) work in various ways.

Our work shares some features with this previous literature in that we also consider a

model in which a supplier must provision capacity to meet an uncertain demand, where 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.

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 smaller 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.

2 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 provider faces a competitive market, the cloud provider then chooses a capacity level Q to maximize efficiency, while setting a price p that results in zero expected profit.

2.1 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 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.

3 Results

3.1 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 will 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 $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$.*

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

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 $r(Q)$, the cloud provider should set a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - r(Q))\sigma(N)$. This implies that to set the efficiency-maximizing level of capacity, the cloud provider should set a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - r(Q))\sigma(N) = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$. □

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 $r(Q) = \frac{c}{V}$ 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 will vary 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.*

Proof. 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 a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$.

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. □

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.*

Proof. 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 capacity 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 a level of capacity $Q = \mu(N) + \Phi^{-1}(1-\frac{c}{V})\sigma(N)$. 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 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 will set 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.

3.2 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.*

Proof. We know from Lemma 1 that for sufficiently large values of N, the cloud provider sets a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$. 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. □

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 have seen 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.*

Proof. 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 would set a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{V})\sigma(N)$. 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. $\square$

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. By combining this result with the result in Theorem 3, it follows that it is more efficient to place excess demand in larger regions than in smaller regions.

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.

3.3 Provisioning Capacity to Decrease Stockout Probabilities

Suppose circumstances adjust in such a way that a cloud provider wants to provision more capacity to decrease the probability of a stockout (*i.e.* the probability there will not be enough capacity to meet all customer demand) in a region. This might happen if, for example, customers place more value V on being able to obtain compute when they want it. It could also arise if the cost c for providing a unit of compute declines. Since we have seen in the proof of Lemma 1 that the cloud provider should choose a level of capacity Q in a given region in such a way that the probability of a stockout in the data center, $r(Q)$, satisfies $r(Q) = \frac{c}{V}$, it follows that if either c declines or V increases, the optimal amount of capacity Q to provision in a region will increase as well.

How would the incremental amount of capacity that a cloud provider provisions in order to target a lower probability of a stockout vary with the size of the region? We address this question in the following theorem:

Theorem 5 *The amount of additional capacity a cloud provider would have to provision in a region in order to decrease the probability of a stockout from r to some $r' < r$ is (i)*

increasing in the number of customers in the region N and (ii) decreasing on a percentage basis in the number of customers in the region N.

Proof. We know that for sufficiently large N, the distribution of total demand, $D = \sum_{i=1}^{N} D_i$, is drawn from the distribution $\Phi(D|\mu(N), \sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$. A consequence of this is that if a cloud provider wishes to ensure that the probability of a stockout in a region is r, then it is necessary to provision a total of $Q = \mu(N) + \Phi^{-1}(1-r)\sigma(N)$ capacity in the region. Thus if a cloud provider wants to provision enough additional capacity in the region to decrease the probability of a stockout from r to some $r' < r$, then the cloud provider would provision an additional $(\Phi^{-1}(1-r') - \Phi^{-1}(1-r))\sigma(N)$ capacity in the region. Since this expression is increasing in N, it follows that the amount of additional capacity a cloud provider would have to provision in a region in order to decrease the probability of a stockout from r to some $r' < r$ is increasing in N.

Now since a cloud provider would provision a total of $Q = \mu(N) + \Phi^{-1}(1-r)\sigma(N)$ capacity in the region if the cloud provider wishes to ensure that the probability of a stockout is r, the fractional increase in capacity in this region if a cloud provider wishes to reduce the probability of a stockout from r to r' would be $\frac{(\Phi^{-1}(1-r')-\Phi^{-1}(1-r))\sigma(N)}{\mu(N)+\Phi^{-1}(1-r)\sigma(N)} = \frac{\Phi^{-1}(1-r')-\Phi^{-1}(1-r)}{\mu(N)/\sigma(N)+\Phi^{-1}(1-r)}$. Since $\frac{\sigma(N)}{\mu(N)}$ is decreasing in N, it follows that $\frac{\mu(N)}{\sigma(N)}$ is increasing in N, and $\frac{\Phi^{-1}(1-r')-\Phi^{-1}(1-r)}{\mu(N)/\sigma(N)+\Phi^{-1}(1-r)}$ is decreasing in N. Thus the amount of additional capacity a cloud provider would have to provision in a region in order to decrease the probability of a stockout from r to some $r' < r$ is decreasing on a percentage basis in N. □

The result in Theorem 5 indicates that if a cloud provider wants to decrease the probability of a stockout in each of its regions, then the absolute difference in the total amount of capacity provisioned in large regions and small regions will increase, but the percentage difference will decline.

4 Empirical Results

This section presents empirical results that quantify the magnitude of one of our theoretical results identified in the previous section. We use data from Microsoft Azure to illustrate the extent to which price varies with the size of the region.

Throughout this section we use data from the wide variety of virtual machines (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.

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 general purpose, compute optimized, and memory optimized VM that was available for purchase to a US customer as of March 2020 (46 regions in total), we noted the price, total supply, and total demand for this type of VM in each region. 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.

The results of this analysis revealed significant negative correlation between the price and the size of the region. For each of the types of VMs in question, we estimated a correlation coefficient between price and region size that fell somewhere between -0.27 and -0.48, with an average correlation coefficient of -0.39. In addition, for each of these types of VMs, we also estimated a correlation coefficient between price and the log of the region size that fell somewhere between -0.33 and -0.43, again with an average correlation coefficient of -0.39. Finally, the average prices for these types of VMs was 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.

5 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 finding on how prices vary with the size of a region is consistent with practice at Microsoft Azure, as prices tend to be $10-20\%$ higher in Azure's smallest regions than in its largest regions.

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] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

[17] 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.

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

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

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