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

Patrick Hummel and Michael Schwarz
Microsoft Corporation

Public Cloud
- Public clouds provide computing capacity which can be rented on-demand for computation
- Major public clouds include Amazon Web Services, Microsoft Azure, and Google Cloud
- Each public cloud has dozens of different regions throughout the world

Regions
- Regions differ in size, price, and utilization rates
- Utilization differences are a key driver of cost differences between regions
- How should firms provision capacity and set prices in different locations?
- Should firms steer customers towards large or small regions?

Business Motivation
- Should internal Microsoft cloud customers be encouraged to use regions with the lowest capacity utilizations?
- Opposite strategy is economically optimal
- Steer customers to regions with high utilization rates (larger regions) can lead to a noticeable cost savings

Notation
- $N$ = number of potential customers in region
- $D_i$ = demand for customer $i$
- $D = \sum_{i=1}^{N} D_i$ = total demand
- $c =$ cost of one unit of compute
- $V =$ customer value for one unit of compute
- $Q =$ amount of compute supplied
- $p =$ price for one unit of compute

Demand Assumptions
- It suffices to assume each $D_i$ is an independent draw from a distribution with bounded support
- The following assumptions are less restrictive:
  - For sufficiently large $N$, $D = \sum_{i=1}^{N} D_i$ 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 satisfies $\Phi(D) = 1 - \Phi(-D)$
  - $\mu(N)$ and $\sigma(N)$ are increasing functions of $N$
  - $\frac{\mu(N)}{\sigma(N)}$ is decreasing in $N$
  - $\sigma(N)$ is a strictly concave function of $N$

Cloud Provider Choices
- The cloud provider chooses capacity $Q$ to maximize efficiency given uncertain demand
- Uncertainty about actual demand is revealed after capacity is chosen
- The cloud provider sets a price $p$ that is increasing in average costs

Optimal Capacity Choices
- **Lemma:** For sufficiently large $N$, the cloud provider sets a level of capacity $Q = \mu(N) + \Phi^{-1}(1 - \frac{c}{p})\sigma(N)$
- The probability that $D > Q$ is then $\frac{c}{p}$

Unfilled Demand
- **Theorem:** For sufficiently large $N$, the expected fraction of demand that will be unfilled by the available capacity is decreasing in $N$
- Result follows from $\frac{\sigma(N)}{\mu(N)}$ being decreasing in $N$

Prices
- **Theorem:** For sufficiently large $N$, the price for a unit of compute is decreasing in $N$
- Excess capacity needed as a fraction of expected demand to be able to meet all customer requests with high probability is lower in larger regions
- Costs and prices are lower in larger regions

Marginal Capacity Costs
- **Theorem:** If $N + 1$ 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 $N$
- $C(N + 1) - C(N) = c(\mu(N+1) - \mu(N)) + \Phi^{-1}(1 - \frac{c}{p})(\sigma(N+1) - \sigma(N))$
- Result follows from concavity of $\Phi(\cdot)$

Empirical Results - Prices
- Considered six types of Azure VMs offered in each Azure region
- Supply by region: capacity for a given type of VM
- Demand by region: demand for a given VM type
- Supply and demand-based measures of region size are nearly perfectly correlated (we use supply)

Empirical Results - Marginal Costs
- Average correlation between price and region size $= -0.43$ (range between $-0.38$ and $-0.48$
- Average log correlation between price and region size $= -0.5$ (range between $-0.37$ and $-0.6$
- Prices in the smallest $\frac{1}{4}$ of regions are $10 - 20\%$ higher than those in the largest $\frac{1}{4}$ of regions (exact price differences vary by VM type)

Empirical Results - Marginal
- For each day $t$ in a one-year period, we calculated the total supply $Q_t$ and the total demand $D_t$ in each region
- Ran a linear regression of $Q_t$ on $D_t$ for each region
- The regression coefficient gives a measure of the ratio of changes in capacity supplied to changes in demand for each region
- Correlation between coefficient and region size $= -0.3$; log correlation $= -0.37$

Hyper-Flexible Customers
- Workloads can be deployed in any region after observing the demand of other customers (not currently offered by cloud providers)
- Small number of hyper-flexible customers $\Rightarrow$ negligible incremental hardware cost
- Large number of hyper-flexible customers $\Rightarrow$ cost disadvantage of small regions vanishes

Conclusion
- The fraction of unfilled demand and prices are lower in larger regions
- Marginal capacity costs are lower in larger regions
- Results are consistent with empirical evidence from Microsoft Azure