In broad terms, cloud computing refers to the provision of computing resources, such as storage, data management and processing, over a network of remote servers hosted on the Internet. Amazon, currently the dominant service provider in this market, offers its computing resources under three different pricing models: pay-as-you-go (on-demand instances), pay-as-you-go under contract (reserved instances), and second price auction (spot instances). Looking at Amazon’s price trajectories for a subset of 546 products during the period July 15, 2014 - January 15, 2015, on average, we find 12% of the time the spot price was higher than the corresponding on-demand price, and for 86% of the products, procuring spot instances, with sufficiently high bids so as never to be shut off, would cost less than on-demand instances for the whole 6-month-period.

Asymptotic analysis of large scale service systems under a priority rule (akin to guaranteed vs. spot) suggests that stochastic fluctuations in the guaranteed capacity utilization and spot demand will be second order in magnitude. Assuming that spot capacity is not vanishingly small, second order fluctuations seem unlikely to explain variations in spot prices observed in the data. The operating cost of capacity may not be constant over time, and indeed may vary as a function of the intertemporal price of electricity, but cross-product analysis does not seem to indicate this as a primary explanation of the spot price fluctuations.

In this paper we study an infinite capacity, revenue maximizing provider of cloud computing resources. The service provider offers a) guaranteed (on-demand instances) and b) best effort (spot instances). The latter are procured via a second price auction. Given that capacity is infinite, there is no competition for resources between the two service classes and amongst the bidders for spot service. Users are heterogeneous with respect to their valuation per hr and disutility per hr of service disruption. The main finding of our work is that under plausible assumptions on the market primitives the service provider may intentionally inject stochastic price fluctuations in the spot service (e.g., by varying her reserve price over time) so as to degrade the spot instance availability for low bidders, incentivizing congestion sensitive and high valuation users to procure guaranteed resources.

Product quality is defined as the fraction of time the product is available to the user. Guaranteed service offers 100% availability. Spot service availability is equal to the fraction of time the user’s bid exceeds the prevailing spot price. The model considers a continuum of customer types that differ with respect to valuation per unit time of service and disutility per unit time of service disruption. Both valuation and sensitivity to congestion are private information. For the majority of the paper we assume customers are one dimensional: value per hr is an affine function of congestion cost per hr of service interruption. This implies that (value/hr)/(cong. sens./hr) ↓ as (cong. sens./hr) ↑.

We solve a deterministic relaxation of the service provider’s revenue maximization problem and show that it is optimal to offer both guaranteed and spot services. The spot option has a potentially stochastic price path, captured by a steady state distribution for the fraction of time spent at different price levels. Users take that description and their own parameters into account when choosing between guaranteed and spot, and how to bid if they select the latter. Perhaps surprisingly, we show that it suffices to fluctuate the spot price between just two price levels, and that the fraction of time that the spot service price is high depends on the constants of the affine relation between congestion cost rate and valuation rate, but not the distribution of types itself.

If (value/hr)/(cong. sens./hr) ↑ as (cong. sens./hr) ↑, we prove that it is never optimal to offer spot service, and the service provider’s problem reduces to finding the marginal user type that will be served. In this setting congestion sensitive users are willing to pay increasingly high amounts, and the service provider focuses exclusively on that segment.

Taking the structure of our derived pricing policies as given, we return to Amazon’s data and estimate the parameters for our model that would best fit the observed price traces. We express the observed spot price path of each product as a multiple of its on-demand price, and find a two price-level CDF that is “closest” to the empirical CDF of the normalized prices. The estimated two-price distribution (per product) can be used to calibrate our model parameters. Results show that the (cross-product) normalized value/hr is between 0.7 and 1.5, the normalized congestion cost/hr is between 0 and 6, and the average % of downtime of the spot service is 0.3. This implies that congestion costs/hr, due to lost revenue and possibly lost goodwill/reputation, can be up to 4x-10x of the value/hr. The realized congestion cost in spot service is of the same order of magnitude as the valuation; for high user types, it is always optimal to choose the guaranteed (always available) service; the lower 1/3 of the user population (in terms of value and congestion cost rates) choose the unreliable spot service.

*This research was supported by a grant from the W. Edwards Deming Center at Columbia Business School.