How to Bid the Cloud – Public Review

Michael Schapira
Hebrew University of Jerusalem
Jerusalem, Israel
schapiram@huji.ac.il

Amazon's Elastic Compute Cloud (EC2) leverages an auction approach, called "spot pricing", to sell cloud resources to users. Spot prices are constantly adjusted by the cloud provider in response to demands, and users' bids that exceed the spot price are accepted. When a user's bid falls below the prevailing spot price, the user's instance is terminated until such a time his bid exceeds the spot price again. Spot pricing thus enables the cloud provider to make efficient use of cloud resources and, at the same time, allows users to enjoy cloud resources at a reduced rate. This auction setting gives rise to two important questions: (1) How should the cloud provider set the spot price? and (2) What is the right bidding strategy for the users?

This paper tackles both of these questions, and makes the following contributions:

- **A model of spot pricing.** Understanding how the cloud provider sets and adjusts spot prices is a challenging task. The paper presents a model for explaining cloud providers' offered spot prices when the provider's goal is to maximize revenue alongside efficiently utilizing capacity.
- **An explanatory model of user bidding.** Selecting the "right" bidding strategy is nontrivial. There is an inherent tradeoff between the risk that the user's job be interrupted and the cost charged for executing the job. High bids naturally entail higher costs for the bidder (if accepted) but reduce the risk that the task be interrupted and the recovery time following an interruption. The authors show how, given a predicted distribution over spot prices (generated by the above model of spot pricing), the optimal bidding strategy for the users can be derived.
- **Case study: parallelized MapReduce.** The paper shows how the proposed bidding strategies can be adapted to parallelized MapReduce jobs. To evaluate the effectiveness of the proposed strategies, the authors performed a variety of experiments on Amazon EC2 by running both single instance-jobs and MapReduce jobs. The results show that using spot instances does reduce the usage cost.

The PC agreed that, given the growing popularity of Infrastructure-as-a-Service in public cloud, the questions studied in this paper are both important and timely. The reviewers appreciated the rigorous game-theoretic/economic approach taken in the paper and the authors' efforts to validate the theoretical findings presented in the paper with real experiments on Amazon EC2.

While some aspects of the model for how spot prices are set certainly do not accurately capture reality, the paper does a good job at explaining these limitations and the PC thought that the insights gained through the mathematical modeling and analyses presented in the paper are valuable. We believe that further research should focus on applying this general approach to more accurate models of spot pricing, including (1) models for setting spot prices that are not only based on revenue maximization and increasing networking utilization, but also reflect other factors such as failure resiliency and minimizing overload, (2) incorporating "soft deadlines", i.e., users whose goal is completing a job before time T with probability at least p for some T and p, and (3) more realistic assignment of values to the parameters in the proposed model. In particular, the model should be contrasted with other models for setting spot prices, e.g., the model in [Agmon Ben-Yehuda et al., "Deconstructing Amazon EC2 Spot Instance Pricing", ACM Trans. Economics and Computation]. Another interesting direction for future research is extending the kind of theoretical and empirical analysis presented in the paper beyond the EC2 case study.

In summary, the PC thought that the paper presents a valuable approach to reasoning about spot pricing and, more generally, IaaS. We hope and believe that this will trigger follow-up work on this topic.