A Report on the Workshop on the Economics of Cloud Computing

NIKHIL R. DEVANUR
Microsoft Research

This is a report on the first Workshop on the Economics of Cloud Computing, which was held in conjunction with the ACM conference on Economics and Computation (EC).

1. INTRODUCTION

The digitization of the world's businesses, and the movement of this digitization into the cloud is akin to the industrial revolution. It is speculated that cloud computing will be to businesses what mobile computing has been to consumers. This raises a whole slew of questions in economics, most of which are deeply entangled with computer science topics. A half-day workshop on the economic aspets of cloud computing was held in conjunction with the ACM conference on Economics and Computation (EC) 2016 in Maastricht. The goal of the workshop was to be the premier platform to raise the most important research questions, to announce the latest results, to exchange ideas, to learn and to get feedback on the state-of-the-art research in this area. The topics of interest for this workshop were broadly set out to be as follows.

Moving to the Cloud. How are current businesses impacted by moving to a cloud enabled world?

New Markets. What new markets emerge as a result of a cloud enabled world? What new economic models come into play?

Cloud Pricing. What are the different pricing or auction mechanisms to sell cloud computing resources, and the pros and cons of each?

Cloud provisioning. What are best practices in the process of provisioning all the requirements for building a datacenter? What economies of scale can be exploited in running large data centers?

Fair allocation. How should one allocate cloud resources in a fair manner in a shared multi-tenant system?

There were 2 keynote speakers, Noam Nisan and Simon Wilkie, and 6 contributed talks. A call for papers for the contributed talks was circulated, and the following program committee decided which papers are to be presented.

—Nikhil R. Devanur (Program Chair)
—Eric Friedman
—Preston McAfee
—Noam Nisan

Author's address : nikdev@microsoft.com

—Eva Tardos

—Adam Wierman

All the details about the workshop can be found at `http://wecc.azurewebsites.net/`. We now present short descriptions of the keynote and the contributed talks.

2. TALKS

2.1 Keynotes

Noam Nisan:ERA: A Framework for Economic Resource Allocation for the Cloud. Noam Nisan from the Hebrew University at Jerusalem opened the talk with his definition of the cloud as a shared computational resource, which is typically a virtual machine at a remote data center. A well designed cloud system should make the most efficient use of the shared resources. For instance, flexible jobs should run during low congestion times, and the most "valuable" jobs must be run during periods of over demand. Simple schemes such as pay-as-you-go and dedicated hosts have obvious inefficiencies. Overcoming these shortcomings requires skills from various disciplines such as computer systems, algorithms, and economics.

The Economic Resource Allocation (ERA) project is a prototype system that is meant to expose cloud design issues at the intersection of all these three disciplines. It is a system for scheduling, reserving and pricing cloud resources. It provides friendly APIs for two interfaces: one user-facing that accepts requests for reservations, which are either accepted at a given price, or rejected; the other interfaces with an existing cloud system and provides it with jobs to run at any point of time. In between these two interfaces sits the ERA algorithm. The system allows plugging and playing with different algorithms, making it easy to compare and contrast them.

The prototype was used to provide a proof of concept for the benefits of combining insights from systems, algorithms, and economics. It showed how a simple economically aware algorithm can significantly improve the efficiency of the system. It provides a unified simulator and a platform over various cloud systems on which to test algorithms; this is a useful tool for future research. The main algorithmic challenges are in predicting future job requests from data, and in making optimal scheduling decisions from these predictions.

Simon Wilkie: The Price of Privacy in the Cloud, or The Economic Consequences of Mr. Snowden. Simon Wilkie from Microsoft Research spoke about estimating the effect of the Snowden revelations on cloud adoption for US based cloud providers. The adoption of the cloud among businesses has been on an upward trajectory ever since the inception of large scale cloud providers in the late 2000s. And then, Snowden's revelations about NSA's spying program in 2013 made consumers who cared about the privacy of their data wary of adopting US cloud providers. What was the effect of this? How much revenue was lost?

Simon and his co-author (Hyojin Song, Microsoft Research) find answers to these intriguing questions by using a global panel dataset of cloud revenues. They build a behavioral model for cloud adoption from the data, and use the non-US based providers as the "control". This then lets them estimate the magnitude of the negative demand shock on US providers due to Snowden. They estimated that

the growth rate of US providers decreased by about 11%; this equalled about 18 billion USD in lost revenues. The US providers reacted to this decreased demand by reducing prices, which led to a "price war". An interesting side effect of this price war was that the market share of US providers eventually went up.

2.2 Contributed Talks

Cloud Pricing: The Spot Market Strikes Back. The decision on which model to use for selling cloud resources is a very real and important one for the providers. [Dierks and Seuken 2016] consider whether offering both a fixed price and a dynamic (spot) price can increase profits over offering either of these alone. Previous work by [Abhishek et al. 2012] showed that the answer is that it doesn't, but based on an assumption that the provider has access to an infinite pool of resources. This paper considers the cost of procuring resources, and shows that a hybrid model can indeed be better for profit. The demand is modeled as a stochastic process similar to the queueing theory models. The system is assumed to be at an equilibrium where the supply (the number of servers provisioned) is equal to the demand. (The expected waiting time of a job is below some threshold.) In the current model, the idle instances of the fixed price market cannot be sold on the spot market. Utilizing such idle instances is one of the main attractions of the spot market, and incorporating this into the model seems an important step.

On-Demand or Spot? Selling the Cloud to Risk-Averse Customers. [Hoy et al. 2016] consider essentially the same quesion as the previous talk, but focus on a risk averse model of a consumer. A concave curve determines the utility of a consumer as a function of her surplus. A dual market with both fixed and spot prices works as follows: bidders first decide whether to reserve an instance using the fixed price market. The available supply is then sampled from a given distribution. Any excess supply that remains after allocating all the reserved instances is then sold through an auction resulting in a spot price. They show how this model explains the existence of such a dual market by showing increased revenue/welfare/efficiency compared to markets with a single option.

An alternate direction to tackle this issue of two markets versus one is to consider time sensitivity of consumers. Consumers whose jobs are time sensitive tend to opt for a reservation market while others would prefer the lower prices in a spot market. Another interesting question is how the conclusion is affected by the presence of competitors who sell imperfect substitutes.

Approximately Efficient Cost Sharing via Double Auctions. [Fischer et al. 2016] propose jointly solving the problems of pricing and procurement for cloud resources. This makes sense since any reasonable objective (welfare/revenue) depends on both the demand as well as the cost. Solving each of them separately (assuming the other as fixed), as is done currently, can be sub optimal. They model this as a cost sharing problem and consider a twist to the standard guarantees by allowing an additive as well as a multiplicative term for approximating efficiency. They present a mechanism that is inspired by the double auction of [McAfee 1992]. In a large market setting, this mechanism attains strategyproofness, budget balance and approximate (additive + multiplicative) efficiency, thus bypassing previous impossibility reults. The result is interesting more generally for other cost sharing

problems as well.

Pretium: Dynamic Pricing and Traffic Engineering for Timely Inter-Datacenter Transfers. [Jalaparti et al. 2016] consider pricing schemes for data transfers between data centers. The current standard practice for such transfers is to have a fixed price per unit of data, but this is inefficient due to the large temporal variation in the requests. Moreover, it is shown via a survey that customers are quite receptive to the idea of time-of-use pricing and trading off the timing of their transfers for the cost of transfers. The paper shows that a dynamic pricing combined with traffic engineering can significantly increase the efficiency of these systems. The methodology is empirical: they uses traces of data transfer from a large data center and replay them under the different schemes. The prices are calculated using the requests from a reference time window from the past; the optimal dual variables of a welfare maximizing linear program give market clearing prices. The process is also shown to limit the users from gaming the system by showing that certain types of gaming don't help (both theoretically and empirically). This paper is an excellent demonstration that simple economic insights can have significant impact on real systems.

Congestion Games with Mixed Objectives. In allocating a shared network bandwidth among many users, two different objectives have beens studied: latency, and bandwidth. In any large data center there are heterogeneous users among whom some care more about latency (video gaming) while others more about bandwidth (media streaming). (While latency is typically additive, bandwidth is a min or a max.) [Feldotto et al. 2016] consider congestion games where the users have different utility functions of these two cost measures. They show that when the agent preferences satisfy a certain monotonicity assumption, a pure Nash equilibrium always exists. Moreover, a lazy best response dynamics converges to it. In the absence of this assumption, pure Nash may not exist and finding one is NP-Hard. Best response dynamics may cycle. An interesting direction for future research would be to incorporate uncertainties, in preferences as well as realized costs.

Dynamic Games for Market Dominance in the Cloud. Traditionally, providing an online service involved a big fixed cost for setting up the IT infrastructure and almost zero marginal cost thereafter. Cloud computing flipped this cost structure so that it is almost all marginal cost and no fixed cost. Standard competitive models in economics suggest that this lowers the cost of entry and hence lead to the market being shared by many competitors. Often, one sees the market in the technology sector being dominated by one or two companies, contrary to this prediction. [Conley 2016] seeks to explain this discrepancy by considering the presence of venture capital funding. The strategies for a venture capital funded firm and a publicly traded firm differ in that the venture capital funded firm can be way more aggressive in spending money on advertisements and promotions than a publicly traded firm. This is because once a venture capitalist commits to a funding, it becomes rational for the company to spend as much of it as possible to gain market share, whereas a similar strategy would be irrational for a publicly traded firm.

The paper introduces an interesting model of competition among firms, and there

is an opportunity to extend it further to capture other aspects of the real markets. For instance, while there are a large number of firms entering the market, there are fewer firms competing for consumers ex post because of either network effects resulting in a winner-take-all situation, or because their exit strategy is to get acquired by a bigger firm.

3. CONCLUSION

With cloud computing fast becoming the de facto way for businesses to handle their IT infrastructure, we expect research into the economic aspects of cloud computing to grow. Given sufficient interest from the community, the workshop could become an annual or a biennial fixture at EC, alongside similar workshops such as the one on ad auctions.

REFERENCES

Abhishek, V., Kash, I. A., and Key, P. 2012. Fixed and market pricing for cloud services. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*. IEEE, 157–162.

Conley, J. 2016. Dynamic games for market dominance in the cloud.

Dierks, L. and Seuken, S. 2016. Cloud pricing: The spot market strikes back.

Feldotto, M., Leder, L., and Skopalik, A. 2016. Congestion games with mixed objectives. In *International Conference on Combinatorial Optimization and Applications*. Springer, 655–669.

Fischer, F., Kash, I. A., Key, P., and Wang, J. 2016. Approximately efficient cost sharing via double auctions.

Hoy, D., Immorlica, N., and Lucier, B. 2016. On-demand or spot? selling the cloud to risk-averse customers. In *International Conference on Web and Internet Economics*. Springer, 73–86.

Jalaparti, V., Bliznets, I., Kandula, S., Lucier, B., and Menache, I. 2016. Dynamic pricing and traffic engineering for timely inter-datacenter transfers. In *Proceedings of the 2016 Conference on ACM SIGCOMM 2016 Conference*. SIGCOMM '16. ACM, New York, NY, USA, 73–86.

McAfee, R. P. 1992. A dominant strategy double auction. *Journal of economic Theory 56,* 2, 434–450.