Pricing Schemes in Cloud Computing: Utilization-Based versus Reservation-Based

In recent years, cloud computing has been recognized as a rising trend in the business world. According to 2014 survey results by Oxford Economics and SAP (2014), more and more companies would become moderate-to-heavy users of cloud services in the subsequent several years. Given this trend, cloud service providers face a tremendous growth opportunity. Industry-wide, revenue of the entire cloud service market reached $110 billion in 2015 or a 28% growth rate on an annualized basis (Synergy Research Group 2016).

However, cloud service providers also face serious challenges, especially the intense competition for customers. Price competition is one of the most significant factors affecting cloud service providers' revenues and profits, and “pricing wars” among major public cloud computing providers, such as Amazon AWS, Microsoft Azure, and Google GCP, have made the headlines in recent years. When a customer decides about whether to choose a cloud provider, factors other than price will also be considered, forming the customer's inherent preference for the provider. Although the customer may prefer a provider, the customer's estimation of the expected cost can act as a deterrent to choosing that provider. The customer will thus make a final decision by comparing the preference to the cost. The calculation of a customer's expected cost under different pricing schemes offered by the competing service providers is thus important.

This research is motivated by our observations of different pricing schemes offered by the leading cloud providers including Amazon, Microsoft, and Google. For the short-term customers, all cloud providers offer the so-called on-demand instances or on-demand machine types; an instance corresponds to a specific virtual-machine configuration in terms of the number of virtual CPU (vCPU), the amount of memory, and the operating system. Typically, a provider uses a fixed price (rate) for each type of on-demand instance. Moreover, due to the intense price competition
for the short-term customers, the fixed rate of similar type of on-demand instances of different service providers tends to converge, which we refer to as the base price (rate) for a type of instance. In contrast, to attract long-term customers, cloud providers have offered various incentive pricing schemes in addition to the fixed rate for a type of instance. We have observed two types of such incentive pricing schemes: one is offered by Amazon or Microsoft as the so-called reserved-instances prices, and the other is offered by Google as the sustained-use discounts.

Under Amazon's reserved-instance pricing scheme, a customer reserves a certain number of instances and pays for those instances during the contract term (e.g., one year) even if sometimes those reserved instances were not used. In return, the customer receives a discount for the reserved instances. In contrast, under Google's sustained-use-discount pricing scheme, a customer does not need to make any decisions but simply deploys the virtual machines based on the realized demand, like what the customer would do when using the short-term, on-demand instances. However, the pricing scheme is designed in such a way that customers who achieve a sustained usage status of each deployed virtual machine would be rewarded with a discount.

In this study, we build a duopoly model where each service provider faces a choice between two pricing schemes: one is referred to as the reservation-based pricing scheme, and the other is referred to as the utilization-based pricing scheme. We answer the following two major questions:

1. How do different types of customers of cloud services choose between the two incentive pricing schemes? How would the average and volatility of a customer's demand for computing capacity affect the customer's choice between the two pricing schemes?
2. How do cloud service providers choose their pricing schemes and optimally set the parameters of the chosen pricing schemes? What is the impact of those service providers' choices of pricing schemes on their profits, revenues, and market shares?
This study makes three main contributions.

Our first contribution is providing an analytical model of the current cloud computing pricing schemes and gaining insights into how such pricing schemes could affect the customers' choices and the key performance measures of the service providers.

Second, our analysis could facilitate the customers' cost estimation and enhance the providers' understanding of the impact of their pricing decisions. We find that a customer's expected per-period cost can be expressed as an effective price times the expected per-period demand; moreover, the effective price is an increasing function of the customer's demand volatility. However, when two competing providers employ different pricing schemes, the two pricing schemes would not be equally favorable to all customer segments: indeed, the reservation-based scheme offered by AWS is more favorable to those who have smaller demand volatility, while the utilization-based scheme offered by GCP is more favorable to those who have larger demand volatility.

Third, our analysis provides insights into the service providers' optimal choices of pricing schemes. We find important market conditions and operational characteristics of the service providers, under which one provider would optimally use the reservation-based scheme, while another provider would optimally use the utilization-based scheme, like the current situation where AWS adopts the reservation-based while GPC adopts the utilization-based scheme. We also examine how the evolving market conditions would change the providers' optimal decisions.

References:
