Zhi-Hong Jian
R09323031
R09323031@ntu.edu.tw

Introduction to
'Efficient Capacity Provisioning for Firms with Multiple Locations: The Case of the Public Cloud'
Patrick Hummel, Michael Schwarz (2020)

2020-09-24

1 What are the questions?

The need of cloud computing has been growing during this accelerating era. The cloud computing market consists of customers who need online computational power and providers such as Google Cloud, Amazon Web Services, and Microsoft Azure. For the cloud providers with multiple locations, they have the freedom to decide the price policy and how much cloud capacity to provision in different regions. This paper, written by Michael Schwarz and Patrick Hummel, who are the top economists on Microsoft, tried to answer the following questions:

- In regions of different sizes, what price should the cloud provider charge?
- In regions of different sizes, how much capacity should the firm provision?
- Suppose there are some customers have the flexibility to be placed in any regions. To maximize the efficiency, what size of region should the cloud provider direct them to ?
- Due to the uncertain demand, it's possible to run out of capacity in some regions. How could the providers do to decrease stock out probabilities?

2 Why should we care about it?

Despite the fact that this paper eyes on the public cloud computing market, the model and the conclusions here could be applied to the firms selling homogeneous product in different regions.

3 How did the authors get there?

They set up a model[1] of a competitive market, which consists of firms with multiple locations and customers with uncertain demand, which follow the cumulative distribution function. Moreover, they also used the data from Microsoft Azure, giving some empirical evidences to their model.

4 What are the authors' answers?

To answer the questions above, the authors derived some important theorems from the model. Let N denotes the numbers of potential customers:

- As N gets larger, the expected fraction of demand that will be unfilled by the available capacity decreases.
- As N gets larger, the price of a unit of compute decreases.
- The firm should encourage those flexible customers to the larger region, where the unit cost for the firms is smaller.
- To decrease the probability of stock out, the firms should increase the number of customers and decrease percentage basis in the number of customers in the region N.
- Empirically, there is an negative correlation between price and region size.

[1] Some important notations and assumption of this model are listed on the back page

5 Real World Example

Actually, what the authors discussed in this paper is a real world example.
Besides public cloud, other companies that provide homogeneous service online, such as platforms like HeroKu and Nvidia cloud gaming service, could also use this model to get some ideas about price and the allocation of their service.

6 Assumptions and Notations of the Model

- N: the number of potential potential users
- D_i: the demand of the customer i.
- Due to uncertainty, they follow the cumulative distribution

$$G_N(D_1, ..., D_N) = \Phi(D|\mu(N), \sigma(N)) = \Phi(\frac{D-\mu(N)}{\sigma(N)})$$

  where $\frac{\sigma(N)}{\mu(N)}$ is decreasing in N and $\sigma(N)$ is a strictly concave function of N.
- Because of the competitive market, the cloud provider chooses a capacity level Q and price p to maximize efficiency, and results in zero profit.
- ***Lemma 1****: For sufficiently large values of N, the cloud provider sets a level of capacity:*

$$Q = \mu(N) + \Phi^{-1}(1-\frac{c}{V})\sigma(N)$$