Stochastic Cost Optimization of 2-Stage Resource Provisioning

Konatham Sumalatha . Anbarasi M S

Department of Computer Science and Engineering, Pondicherry Engineering College. Puducherry, India
Department of Information Technology, Pondicherry Engineering College, Puducherry, India

Received: 19 August 2022 / Revised: 27 August 2022 / Accepted: 19 September 2022

Abstract — As Cloud Computing is a pay-as-you-go model, it offers huge computational resources available for users. Due to uncertainty of demand price in future. Cloud Consumer-facing difficulty to get these resources in a cost-effective & robust manner. In addition to that, heterogeneity in pricing plans is another complexity. Hence, to address this problem we are considering two-stage stochastic programming with 3 pricing models, viz Reservation, On-demand & Spot pricing. This model gives optimal solutions compared to other existing models.

Keywords — Cloud computing, stochastic optimization, resource provisioning, pot pricing

I. INTRODUCTION

Resource provisioning issues in the cloud computing environment can be seen from different views, like Cloud Provider of Infrastructure as a Service (IaaS), Software as a Service (SaaS), and the Cloud Consumer. Each entity attempts to amplify its benefit in the resource provisioning plan stage. Because of various objectives, prerequisites, and imperatives, the resource provisioning issue in cloud computing should be tended to independently, for every entity. In this paper, we consider this problem from the consumer perspective. The end client is an individual or an association, intending to lease computational assets from an open cloud supplier. In cloud computing, provisioning of resources must be decided by the cloud consumer, which is another and complicated task for such users, who are acquainted with working with a fixed arrangement of assets they own. Instead, they experience an entangled dynamic issue of picking the most reasonable sort and number of VMs with the best evaluating plans, for running their application. There are likewise dubious parameters that make this streamlining issue significantly more convoluted. Since the resource request is profoundly dynamic, its example can't be known, or even be precisely anticipated, ahead of time. Besides the cost of assets changes and isn't effectively unsurprising. Various cloud suppliers offer different VM types, to be a specific reservation, on-demand, and spot pricing.

The unit cost of reserved VMs is regularly the most reduced, yet under-provisioning can happen when the held resources can't completely meet the demand. In any case, this issue can be comprehended by provisioning more VMs with either on-demand or spot plans, although the user might be charged a more significant cost. Another potential issue is the over-provisioning issue when reserved VMs are not fully utilized. The cloud consumer-facing the critical issue of managing the above issues and how to optimize the cost.

Although there is much research conducted on resource provisioning from the IaaS [1] and

SaaS [2] cloud provider's view. But in cloud consumers' view, much research is needed as it is in the real world. The paper relates to the OCRP algorithm for the optimization of provisioning cost in cloud computing from the consumer perspective [3]. Demand and price uncertainty are considered to make an optimal selection of resources. To optimize the cost multistage stochastic programming is proposed. However, VMs spot pricing and heterogeneity are lagged in this work. In [4], used two-phase method of Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to resolve the issue from the cloud consumer's view. But, not consider the heterogeneity of resources and different pricing plans. The focus of [6], cost and speed tradeoff to process a lot of autonomous jobs from cloud consumer's side, here it only considered on-demanding model. In [5], based on the future prediction of workload, varied the allocation of VMs to services.

Much existing research above focuses on perfect foresight with deterministic formulations fixed information [7]. However, few [3] considered the uncertainty of demand and price with only one aspect (e.g., real-time pricing), [8] considered simple and artificial data. Most of the above works ignored Pricing and VM heterogeneity.

In this paper, we propose two-stage stochastic programming to optimize cost by opting for the best pricing plans and VM types, demand uncertainty of end-user applications, and provider's price uncertainty (on-demand and spot). We are implementing this mechanism in cloud-sim with a MIP solver. The experimental results prove that our mechanism reduces the operational cost, in contrast to existing works.

The rest of the paper is organized as follows. The problem description and formulation are discussed in section II, then in section III is the experimental study, and in section IV results. Section V, conclusions, and future work.

II. MATHEMATICAL MODEL AND PROBLEM FORMULATION

A. Assumptions

Many cloud providers, provide resources to cloud consumers on a rental basis. Here, we consider a public cloud provider that provides VMs with less cost for running consumers' applications. One such optimal provider in the cloud market is Amazon[9]. Amazon EC2 offers various IaaS solutions. In this work, we assume that cloud consumers rent VMs from Amazon EC2. Since Amazon EC2 instances are widely used IaaS providers in Industry & Academia. Amazon EC2 instance.

Complicated parameters for users while decision-making about Amazon EC2 instances are pricing plans, purchasing variants, and heterogeneity in resource configuration. Pricing plans are reservation pricing, reserving VMs for a specific period (like 1 year,6 months, etc.). This pricing plan charges less usage cost. And also it includes 3 purchasing variants like "all-upfront", which is onetime payment, "partial upfront", which provides monthly payment, and "no upfront". Where the "all-upfront" variant gives more discount compared to other variants.

On-demand pricing allows the consumers to pay per user per hour with no upfront payment and long-term commitment. Cloud consumers may increase or decrease the compute capacity based on the requirement of user application and pay hourly for the used instances. Despite flexibility and reliability, it charges more compared to other pricing plans. The cloud provider may change the price as it is not fixed.

Spot pricing allows the consumers to bid on unused Amazon EC2 instances which fluctuate continuously based on the supply and demand. The bidding strategy here is cloud consumers place the request with instance type and their maximum price pay per hour per instance. The requested instances are allotted if the bid price is greater than or equal to the current spot price. Spot instances are used till the termination by the consumer or its price increases more than the current spot price. The cloud provider shut off the allotted spot instances if the price exceeds the current spot price with a few

minutes' intimations. Spot pricing is applicable for fault-tolerant applications. A big challenge is selecting the best bidding strategy. The most appropriate method for bidding [10] is always to bid lower and send the excess to on-demand price.

The main issue is cloud consumer-facing the complexity of all the above parameters like different pricing plans, purchasing variants, and resource types, and also on-demand and price uncertainty. As reserved resources are more economical users must decide the required VMs in advance without knowing actual need and workload. Hence, cloud consumer needs more information about future workload. To overcome under-provisioning and over-provisioning, an accurate decision can be made at the initial stage. If these resources are not enough in the working phase, then go for on-demand or spot instances through the auction method to fulfill the need. Hence, the problem is divided into two phases: before knowing uncertain parameters decide the number of reserved VMs and then fulfill under-provisioning with on-demand or spot instances.

As the problem involves uncertainty, stochastic programming [11] which is most appropriate is used to solve the problem. Stochastic models are taking advantage of the fact that the probability distributions of uncertain data can be estimated. These two-stage stochastic models are the most widely used models to solve problems that involve uncertainty.

B. Problem Formulation

Minimizing the total cost of provisioning resources the main objective of this problem is. The decision variables are represented with *y*, which indicates the number of each VM type provisioned with variant purchasing variants and pricing plans. See Table 1 for notation. The decision variable y_{ik}^{R} is the number of reserved VM type *i*, subscribed to purchasing variant *k* in the first stage, while y_{ik}^{O} denotes the number of operating VM type *i* with purchasing variant *k* in the second stage. The operating cost is the hourly rate of the reserved VM's utilization cost. Also decision variables y_i^D, y_i^S are the number of on-demand and spot VMs.

Table 1

Notation	Description
I	Set of VM Types
R	Set VM resources (CPU,Memory)
T	Set of Tasks
Cap_i^{CPU}	CPU capacity of VM type i
Cap_i^{memory}	Memory capacity of VM type i
Req_t^{CPU}	Required CPU for the completing task t
Req_t^{Memory}	Required CPU for the completing task t
S	Set of scenarios
K	Set purchasing variants, all-upfront, partial-upfront and no-upfront
MaxBudget	Consumer's maximum budget
C_{ik}^{R}	Reservation cost of VM type i with purchasing variant k
y_{ik}^{R}	Number of reserved VMs type i with purchasing variant k
C_{ik}^{O}	The operational cost of VM type i with purchasing variant k
y_{ik}^{O}	Number of operational VM type i with purchasing variant k
C_i^D	On-demand cost of VM type i
y_i^D	Number on-demand VMs of type i
C_i^S	The spot cost of VM type i
y_i^S	Number spot VMs of type i

Cost functions are calculated as shown below

- The total Reservation Cost, reservation cost of total VMs at the first stage.

$$C_{ik}^{R} = \sum_{i \in I} \sum_{k \in K} y_{ik}^{R}\, c_{ik}^{R} \dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

- The total expending (operational) Cost, the actual cost of resources at the working stage.

$$C_{ik}^{O} = \sum_{i \in I} \sum_{k \in K} y_{ik}^{O}\, c_{ik}^{O} \dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

- The total on-demand cost, the actual cost of on-demand resources.

$$C_i^D = \sum_{i \in I} y_i^D \, c_i^D \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

- The total cost of spot instances.

$$C_i^S = \sum_{i \in I} y_i^S \, c_i^S \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The objective function is to minimize the total cost of provisioning resources.

$$\min \mathrm{Z} = \sum_{i \in I} \sum_{k \in K} y_{ik}^R c_{ik}^R + IE[\phi(y_{ik}^R, \omega)] \ldots\ldots\ldots\ldots\ldots (5)$$

subject to:

$$y_{ik}^R \in \mathrm{N} - - - - - - - - - (5.1)$$

Where, $IE[\phi(y_{ik}^R, \omega)]$ is expected cost at the second stage for scenario ω. And ϕ is the resource function at expending stage, which can be shown as

$$Minimize \sum_{\omega} \left(yc_{ik}^O + \mathrm{y}c_i^D + y_i^S c_i^S \right) \ldots\ldots\ldots\ldots\ldots (6)$$

subject to:

$$y_{ik}^O \leq y_{ik}^R - - - - - - - - - - - - - - - - (6.1)$$

$$z \leq MaxBudget - - - - - - - - - - - (6.2)$$

$$TotalCPU \geq \sum_{t \in T} Req_t^{CPU} - - - - - - (6.3)$$

$$TotalMemory \geq \sum_{t \in T} Req_t^{Memory} - - - (6.4)$$

The constraint (6.1) shows that the number of operational VMs must not exceed the number of reserved VMs. Constraint (6.2) shows that the optimized cost obtained should not be more than the consumer's maximum budget. And constraints (6.3) and (6.4) show that total CPU and memory is greater than or equal to the required amount of CPU and memory.

III. EXPERIMENTAL STUDY

We are considering VM utilization and task length for the workflows taken from Google cluster traces which are relevant to our work [13]. Probably 60% of jobs in the cloud complete their execution within 15 minutes and about 85% of job lengths are less than 50 to 60 minutes [12]. In addition, the interactive and real-time tasks require less amount of CPU and memory when compared to scientific batch jobs [12].

Hence, the proposed method is evaluating with workload traces of the Google cluster traces [14] which is most relevant for our work. These traces consist of two lakh small jobs with a short run-time period. These data are real users' request logs from Google for the tasks like online streaming, web search, translation works, etc. The main characteristics of the data used in this work include total CPUs allotted, Total Memory utilized, User ID, and Group ID fields. 12 user groups with different CPU and Memory usage patterns are considered. Use 12 users' groups as our problem is solved from the cloud consumer's perspective for evaluation,

In addition to the above data, we are taking into consideration the VM prices for all three pricing plans. The IaaS cloud provider's VM prices from Amazon EC2 [15], i.e., the standard reserved, operating, on-demand, and spot prices [16] over May 2016

A. Cloud sim toolkit

Cloud computing is a pay-as-you-use model, which delivers infrastructure (IaaS), platform (PaaS), and software (SaaS) as services to users as per their requirements. Cloud computing exposes data centers' capabilities as network virtual services, which may include the set of required hardware, application with support of the database as well as the user interface. This allows the users to deploy and access applications across the internet which is based on demand and QoS requirements.

This simulation toolkit allows[17] the researchers as well as cloud developers to test the

performance of the potential cloud application for performance testing in a controlled and easy setup environment. And Also allows fine-tuning the overall service performance even before it is deployed in the production environment.

The proposed approach is stochastic programming based so cloud simulator a lot not enough to solve, in addition to that we need one of the optimization solvers called MIP (Mixed Integer Programming) solver.

B. Pricing Alternatives

- ROS(Reservation-On-demand-Spot) pricing: This is the proposed one to get the optimal VMs with additional spot instances.
- RO (Reservation-On Demand) pricing: This includes only reservation and on-demand instances but no spot instances.
- OS (On Demand-Spot) pricing: Considers on-demand and spot without reserved instances.
- O (On-demand) pricing: Includes only on-demand options without reserved and spot instances. It is useful for on and off short-duration workloads.

Each alternative has its benefit based on the workloads.

IV. RESULTS

Comparison among all the pricing alternatives shows in provisioning compared to all other alternatives. Figure 1 shows the graph for the data shown in table 2. On-demand pricing incurs the highest price compared to other approaches. RO and ROS are nearly equal for some groups of workloads.

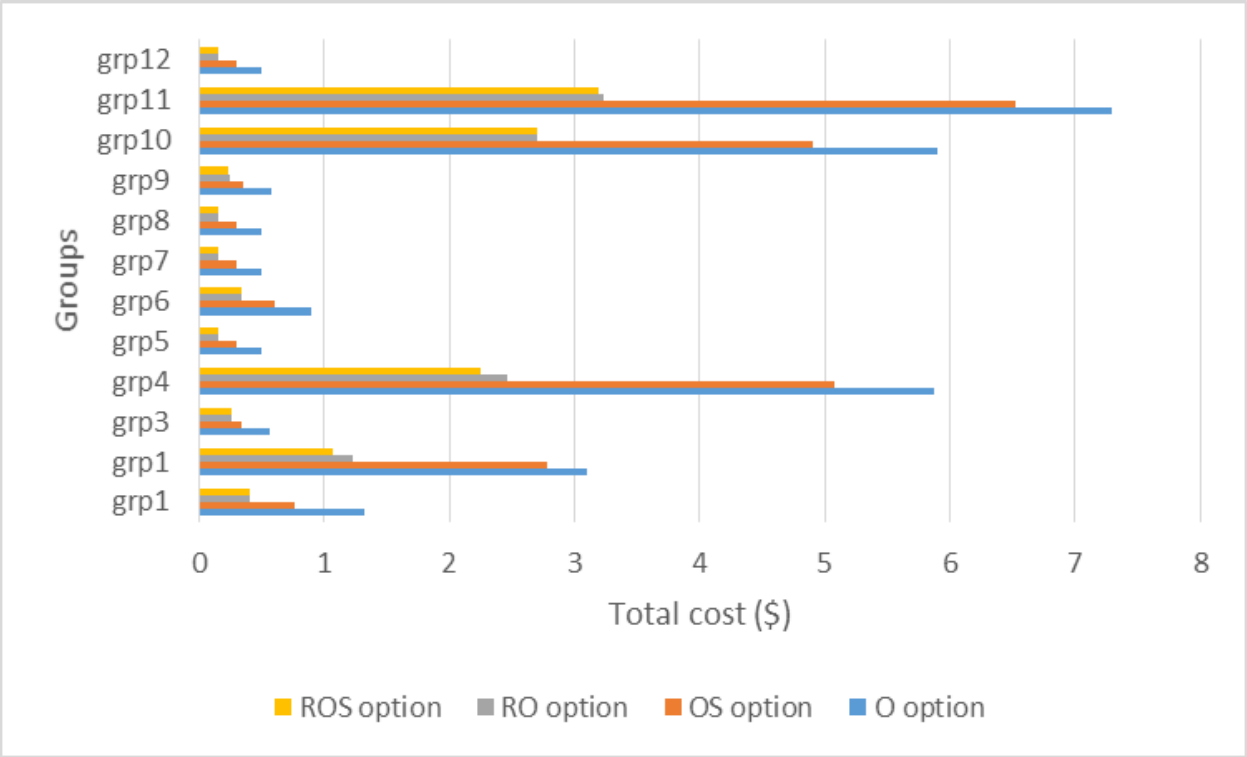

Fig 1. Cost comparison among all alternatives

Table 2: VM Cost calculated for all the alternatives of 12 different groups

	grp1	grp1	grp3	grp4	grp 5	grp6	grp 7	grp8	grp9	grp1 0	grp11	grp1 2

O option	1.32	3.1	0.564	5.874	0.5	0.9	0.5	0.5	0.58	5.9	7.3	0.5
OS option	0.768	2.786	0.348	5.08	0.3	0.605	0.3	0.3	0.36	4.9	6.52	0.3
RO option	0.41	1.231	0.2664	2.4602	0.15	0.34	0.15	0.15	0.2436	2.7	3.2386	0.15
ROS option	0.41	1.074	0.2664	2.2565	0.15	0.34	0.15	0.15	0.2386	2.7	3.1986	0.15

Each group of users has its demand-based workload patterns. For groups 1,3,5,8 the total cost of RO and ROS is equal but for some groups like 2,4,9 and 11, it varies because their workload patterns are different.

Figure 2 shows the number of reserved VMs required in ROS and RO options for 30 runs. For all user groups number of reserved VMs in RO pricing is greater than or equal to the number of VMs in the ROS option.

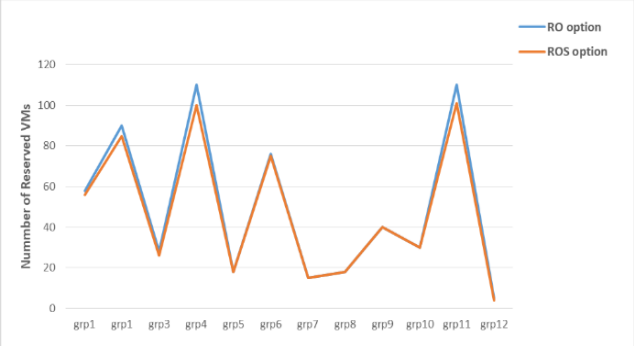

Fig 2. Reserved VMs comparison between RO and ROS

The cloud consumer must decide the optimal number of VMs at the reservation phase i.e first stage to avoid more expensive on-demand resources in the second stage if low-cost VMs are not available. As shown in figure 3, ROS pricing incurs less cost compared to other alternatives and it shows the total cost and the total number of reserved VMs.

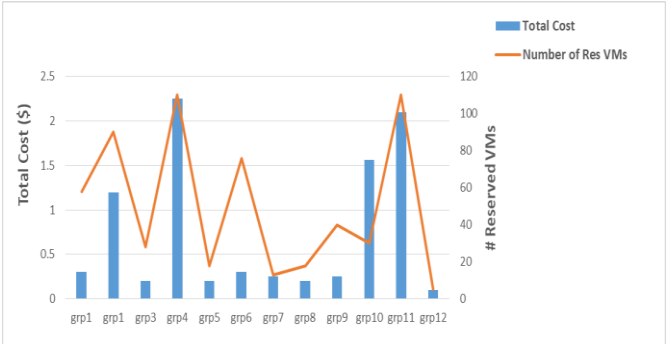

Fig 2. Reserved VMs comparison between RO and ROS

V. CONCLUSION AND FUTURE WORK

Here, cloud resource provisioning is viewed from a consumer point of view. As it includes heterogeneity and uncertainty in VMs and prices, our proposed two-stage stochastic approach determined the optimal number of VMs. The proposed ROS pricing is more economical than other pricing models. Here, in experiment evaluation, we considered user groups with different workload patterns of various CPU patterns. Overall cost should be optimized with the auction approach of spot instances. We plan to extend our work with multiple periods with multistage stochastic programming to get more accurate results.

REFERENCES

1. Chaisiri, S., Lee, B. S., & Niyato, D. (2009, December). Optimal virtual machine placement across multiple cloud providers. In *2009 IEEE Asia-Pacific Services Computing Conference (APSCC)* (pp. 103-110). IEEE.

2. Li, S., Zhou, Y., Jiao, L., Yan, X., Wang, X., & Lyu, M. R. T. (2015). Towards operational cost minimization in hybrid clouds for dynamic resource provisioning with delay-aware optimization. *IEEE Transactions on Services Computing*, *8*(3), 398-409.
3. Chaisiri, S., Lee, B. S., & Niyato, D. (2011). Optimization of resource provisioning cost in cloud computing. *IEEE transactions on services Computing*, *5*(2), 164-177.
4. Adamuthe, A. C., Bhise, V. K., & Thampi, G. T. (2013, November). Solving resource provisioning in cloud using GAs and PSO. In *2013 Nirma University International Conference on Engineering (NUiCONE)* (pp. 1-5). IEEE.
5. Ali-Eldin, A., Kihl, M., Tordsson, J., & Elmroth, E. (2012, June). Efficient provisioning of bursty scientific workloads on the cloud using adaptive elasticity control. In *Proceedings of the 3rd workshop on Scientific Cloud Computing* (pp. 31-40).
6. Genaud, S., & Gossa, J. (2011, July). Cost-wait trade-offs in client-side resource provisioning with elastic clouds. In *2011 IEEE 4th International Conference on Cloud Computing* (pp. 1-8). IEEE.
7. Ahmed, S. T., Ashwini, S., Divya, C., Shetty, M., Anderi, P., & Singh, A. K. (2018). A hybrid and optimized resource scheduling technique using map reduce for larger instruction sets. *International Journal of Engineering & Technology*, *7*(2.33), 843-846.
8. Teng, F., & Magoules, F. (2010, June). Resource pricing and equilibrium allocation policy in cloud computing. In *2010 10th IEEE International Conference on Computer and Information Technology* (pp. 195-202). IEEE.
9. Zafer, M., Song, Y., & Lee, K. W. (2012, June). Optimal bids for spot vms in a cloud for deadline constrained jobs. In *2012 IEEE Fifth International Conference on Cloud Computing* (pp. 75-82). IEEE.
10. Sumalatha, K., & Anbarasi, M. S. (2020, January). Cloud Service Selection Using Fuzzy ANP. In *Annual Convention of the Computer Society of India* (pp. 59-70). Springer, Singapore.
11. Teng, F., & Magoules, F. (2010, June). Resource pricing and equilibrium allocation policy in cloud computing. In *2010 10th IEEE International Conference on Computer and Information Technology* (pp. 195-202). IEEE.
12. Ahmed, S. T., Singh, D. K., Basha, S. M., Nasr, E. A., Kamrani, A. K., & Aboudaif, M. K. (2021). Neural Network Based Mental Depression Identification and Sentiments Classification Technique From Speech Signals: A COVID-19 Focused Pandemic Study. *Frontiers in public health*, *9*.
13. Di, S., Kondo, D., & Cirne, W. (2012, September). Characterization and comparison of cloud versus grid workloads. In *2012 IEEE International Conference on Cluster Computing* (pp. 230-238). IEEE.
14. https://research.google/tools/datasets/google-cluster-workload-traces-2019/.
15. https://aws.amazon.com/pricing/
16. https://aws.amazon.com/ec2/spot/pricing/
17. https://www.cloudsimtutorials.online/cloudsim-simulation-toolkit-an-introduction/