A review on various optimization techniques of resource provisioning in cloud computing

K. Sumalatha¹, M. S. Anbarasi²

¹Department of Computer Science and Engineering, Annamacharaya Institute of Technology and Sciences, India
²Department of Information Technology, Pondicherry Engineering College, India

ABSTRACT

Cloud computing is the provision of IT resources (IaaS) on-demand using a pay as you go model over the internet. It is a broad and deep platform that helps customers build sophisticated, scalable applications. To get the full benefits, research on a wide range of topics is needed. While resource over-provisioning can cost users more than necessary, resource under provisioning hurts the application performance. The cost effectiveness of cloud computing highly depends on how well the customer can optimize the cost of renting resources (VMs) from cloud providers. The issue of resource provisioning optimization from cloud-consumer potential is a complicated optimization issue, which includes much uncertainty parameters. There is a much research avenue available for solving this problem as it is in the real-world. Here, in this paper we provide details about various optimization techniques for resource provisioning.

Keywords:
IaaS
Optimization techniques
Resource provisioning
Uncertainty
Virtual machines

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

1. INTRODUCTION

Cloud computing is a broad and deep platform that helps customers build sophisticated scalable applications. Cloud provisioning is the allocation of a cloud provider's resources to a customer [1]. When a cloud provider accepts a request from a customer, it must create the appropriate number of virtual machines (VMs) [2] and allocate resources to support them. In this context, the term provisioning simply means “to provide”.

The Provisioning has been done in several different ways. a) Advance Provisioning: The customer requests the provider for services and the provider prepares the appropriate resources [2] in advance. The customer is charged a flat fee or is billed on a monthly basis. b) Dynamic Provisioning: The provider allocates more resources when they are needed and removes them [3] when they are not needed. The customer is billed on a pay-per-use basis.

Provisioning allows optimal allocation of resources to consumers in a finite time to achieve desired quality of service. Here, the problem is either the user gets over-provisioning or under-provisioning. Formally, provisioning problem involves uncertainty parameters while choosing resources subject to some constraints to optimize some objective function. The aim is to develop an optimized method that reduces over provisioning and under provisioning problems. It has remained a topic of research in various fields for decades, may it be supply of electricity or water to consumers [4].

In recent years, distributed computing paradigm [5] has gained much attention due to high scalability, reliability, and flexibility. Thus stochastic based techniques deal with these problems by
providing near optimal solutions within the reasonable amount of time. In this paper, we present a review of various optimization techniques based on various parameters.

2. RESOURCE PROVISIONING PRICING MODELS

Reserved instances are cheapest resources. Price is based on the period of subscription (static). Customer must reserve the resources in advance. Customer might overpay for the resources reserved if he/she doesn’t use them extensively and he might under pay for the resources reserved for long time [6].

On-demand instances are the highest priced resources. Price is set by the service provider and remains constant. Consumer has to pay per use. Customer is aware of the exact price to be paid. Resources are reserved for the customer for the paid period of time [7]. Service provider might reserve the resources for longer than the customer’s utilized. Service provider cannot raise the price when demand is high; when demand is low, the user pays higher than the market price.

Spot Instances [8] allow you to specify the maximum hourly price that you are willing to pay to run a particular instance type, usually lower than the On-Demand rate. These are suitable for both customers and the service provider because the price is set according to the level of supply and demand. Less scalability of high demand in the market than fixed pricing, the spot instances are the unused on-demand instances.

The Spot Price fluctuates based on supply and demand for instances, but customers will never pay more than the maximum price they have specified. If the Spot Price moves higher than a customer’s maximum price, the customer’s instance will be shut down by the cloud provider [9]. Figure 1 shows three pricing models.

3. OPTIMIZATION TECHNIQUES

In Computer Science and operation research, Optimization is referred to as the selection of best element from a set of alternatives with regard to some criteria. Two types of entities involved in cloud are cloud service provider and cloud consumer. Cloud service providers provision their resources on rental basis to cloud consumers and cloud consumers submit their requests for provisioning the resources. They both have their own motivations when they become part of cloud environment. Consumers are concerned with the performance of their applications, whereas providers are more interested in efficient utilization of their resources. Thus these Optimization Techniques [10] can be classified into two types: Static techniques and Dynamic techniques. Following are some of the optimization criteria followed while provisioning resources in cloud environment as shown in Figure 2.
3.1. **Static provisioning**

Cloud service provider allocates resources to consumer in advance. That is cloud consumer has to decide how much capacity of resources he/she wants statically means before using them. The deterministic approach, integer programming, linear programming and deadline provisioning algorithm show how resources are allocated statically for deadline based workflows. Here we consider only the expected values for all the parameters. In static provisioning we may get over provisioning or under provisioning problem.

3.2. **Dynamic provisioning**

Cloud service provider allocates resources to consumer as needed and when required and they have to pay per use. This is also called as pay as you go model. Here, genetic algorithm, stochastic approach, approximate dynamic programming, benders decomposition and average approximation are used for provisioning resources dynamically.

3.2.1. **Genetic algorithm**

Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover, and selection. Genetic algorithms do not scale well with complexity, operating on dynamic data sets is difficult.

3.2.2. **Stochastic approach**

In this section, the stochastic programming with multistage recourse is presented as the core formulation. First, the original form of stochastic integer programming formulation is derived. Then, the formulation is transformed into the deterministic equivalent formulation (DEF) which can be solved by traditional optimization solver software.

3.2.3. **Approximate dynamic programming**

Approximate Dynamic Programming (ADP) is a powerful technique to solve large scale discrete time multistage stochastic control processes, i.e., complex Markov Decision Processes (MDPs). These processes consist of a state space $S$, and at each time step $t$, the system is in a particular state $S_t \in S$ from which we can take a decision $x_t$ from the feasible set $X_t$. This decision results in rewards or costs, typically given by $C_t(S_t, x_t)$, and brings us to a new state $S_{t+1}$ with probability $P(S_{t+1}|S_t, x_t)$, i.e., the next state is conditionally independent of all previous states and actions. Therefore, the decision not only determines the direct costs, but also the environment within which future decisions take place, and hence influences the future costs. The goal is to find a policy. A policy $\pi \in \Pi$ can be seen as a decision function $X_t(\pi(S_t))$ that returns a decision $x_t \in X_t$ for all states $S_t \in S$, with $\Pi$ being the set of potential decision functions or policies. The problem of finding the best policy can be written as (1).

$$\min_{\pi \in \Pi} E^\pi \{ \sum_{t=0}^{T} \gamma C_t(S_t, X_t(\pi(S_t))) \} \quad (1)$$

where $\gamma$ is a discount factor, and $T$ denotes the planning horizon, which could be infinite.

3.2.4. **Benders decomposition**

The goal of this algorithm is to break down the optimization problem into multiple smaller problems which can be solved independently and parallel. The Benders decomposition algorithm can decompose integer programming problems with complicating variables into two major problems: master problem and sub problem.

3.2.5. **Sample average approximation**

It may not be efficient to get the solution of the algorithm if the number of scenarios is more by solving the stochastic programming formulation defined in directly if all scenarios in the problem are considered. The sample-average approximation (SAA) approach is used to address this complex problem. This approach is applied on a set of scenarios.

4. **COMPARISON OF RESOURCE PROVISIONING TECHNIQUES**

In [11], proposed an algorithm called optimal cloud resource provisioning algorithm (OCR) to overcome resource under provisioning and over provisioning. Authors applied various optimization techniques to minimize the user's cost. This approach includes an algorithm called optimal cloud resource provisioning algorithm (OCR) to overcome resource under provisioning and over provisioning. Authors applied various optimization techniques to minimize the user's cost with more scenarios. Minimizing both
under-provisioning and over-provisioning problems under the demand and price uncertainty in cloud computing environments is our motivation to explore a resource provisioning strategy for cloud consumers. The following Table 1 shows proposed optimization techniques and their merits and demerits.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Problem Proposed</th>
<th>Optimization Technique Applied</th>
<th>Inputs</th>
<th>Merits</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost &amp; QoS</td>
<td>[12] Resource provisioning and scheduling on IaaS Cloud</td>
<td>Particle Swarm Optimization (PSO) &amp; Meta Heuristic Optimization.</td>
<td>Workflows</td>
<td>Well suited for work-flows with smaller size</td>
<td>It Takes More Cost to meet the deadlines of larger workflows. Implement on cloud environment</td>
</tr>
<tr>
<td></td>
<td>[15] Dynamic Resource Provisioning for Data Streaming Applications</td>
<td>CPU Allocation Algorithm</td>
<td>CPU cycles</td>
<td>Reduce the processing time and budget</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Optimization of Resource Cost</td>
<td>OCRP Algorithm with Stochastic Optimization</td>
<td>Resource requests</td>
<td>Minimize the cost &amp; avoid over &amp; under provisioning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Resource provisioning and scheduling on IaaS Cloud[19]</td>
<td>Particle Swarm Optimization (PSO) &amp; Meta Heuristic Optimization.</td>
<td>Workflows</td>
<td>Well suited for work-flows with smaller size</td>
<td>It Takes More Cost to meet the deadlines of larger workflows.</td>
</tr>
<tr>
<td></td>
<td>[21] Dynamic Provisioning in Hybrid Cloud using Aneka Platform</td>
<td>Map Reduce model, Deadline Priority Provisioning algorithm min-min or max-min algorithm</td>
<td>User requests,</td>
<td>Reduces total execution time</td>
<td>Applicable only for hybrid clouds. Dependencies among tasks can be modeled</td>
</tr>
<tr>
<td></td>
<td>[22] Application of Selective Algorithm for Effective Resource Provisioning</td>
<td>Efficient resource allocation (ERA) Algorithm</td>
<td>File size of user requests</td>
<td>Minimize make span</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[23] An Efficient Architecture and Algorithm for Resource Provisioning in Fog Computing</td>
<td>stochastic scheduling</td>
<td>Sources of uncertainty</td>
<td>Reduces the response time</td>
<td>Applicable for only reserved instances</td>
</tr>
</tbody>
</table>

5. CONCLUSION AND FUTURE WORK
The ultimate goal of cloud computing is to satisfy both cloud consumer and cloud service provider. Broad research is needed to find best optimal provisioning technique. This paper presents various optimized provisioning variants with their merits and challenges while considering different parameters and different algorithms. The efficient dynamic optimization provisioning is one of the primary challenges in cloud environment, because a tradeoff between professional SLA [25] and QoS constraint like max resource...
utilization, cost etc. In future, we will propose a method for Dynamic optimized resource provisioning. It has dynamic nature so that the cost of provisioning resources may be reduced without SLA violations.

REFERENCES
BIOGRAPHIES OF AUTHORS

Ms. K. Sumalatha is currently working as Assistant Professor in Annamacharya Institute of Technology and Sciences and Research Scholar in the Department of Computing Science and Engineering at Pondicherry Engineering College. Her research work focuses on Cloud Computing, Optimization Techniques, and Networks.

Dr. M. S. Anbarasi is currently working as Assistant Professor in the Department of Information Technology at Pondicherry Engineering College. Her area of interest mainly focuses on Data Mining, Databases, and Cloud Computing.