Available online at www.sciencedirect.com
ScienceDirect
ICT Express 3 (2017) 107–110
www.elsevier.com/locate/icte

Cost-optimized configuration of computing instances for large sized cloud systems[☆]

Woo-Chan Kim[a], Ohyun Jo[b,*]

[a] *S/W R&D Center, Samsung Electronics Co., Ltd., Republic of Korea*
[b] *Electronics and Telecommunications Research Institute, Republic of Korea*

Received 23 March 2016; received in revised form 21 September 2016; accepted 13 October 2016
Available online 26 October 2016

Abstract

Cloud computing services are becoming more popular for various reasons which include 'having no need for capital expenditure' and 'the ability to quickly meet business demands'. However, what seems to be an attractive option may become a substantial expenditure as more projects are moved into the cloud. Cloud service companies provide different pricing options to their customers that can potentially lower the customers' spending on the cloud. Choosing the right combination of pricing options can be formulated as a linear mixed integer programming problem, which can be solved using optimization.

Keywords: Cloud system; Cost optimization; Computing instances

1. Introduction

Cloud computing, which enables people to store, manage and process data over the internet, has been receiving a great deal of attention. Many cloud computing services are provided in the forms of IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service) [1]. Additional cloud computing services such as STaaS (STorage as a Service) [2] and MLaaS (Machine Learning as a Service) [3] are also now being offered. Among these services, IaaS is actually the most basic service that provides physical or virtual computing instances and their resources such as storage, IP (Internet Protocol) addresses, and load balancers.

For convenient and effective management, many companies are using the representative IaaS services such as Amazon EC2 (Elastic Compute Cloud), Windows Azure, and Google Compute Engine instead of their own infrastructure. However, as the computing resources required for a company's business or service increase, so does the cost.

We developed a simulation tool for minimizing the price of IaaS services. Our approach obtains an optimized configuration of computing instances based on a variety of IaaS pricing policies. The current version of the simulation tool is built for Amazon EC2, a product of AWS (Amazon Web Services) which has been a champion in cloud computing services since 2006. However, our simulation tool can easily be extended for use with other cloud systems since most IaaS services have similar pricing policies.

Our simulation tool, named Ribon (Reserved Instance simulation tool Based ON R), is open source and available on GitHub.

The reminder of this paper is as follows. Some related works and our contributions are presented in Section 2. Section 3 describes the overall system structure and optimization algorithm, and Section 4 shows the various simulation results. Finally, we present the conclusions of this study in Section 5.

2. Related works

Owing to the popularity of cloud services, there have been a number of studies in the various pricing policies. Several studies use reserved instance to minimize the price of cloud services [4–7].

* Corresponding author.
E-mail address: ohyun.jo@etri.re.kr (O. Jo).
Peer review under responsibility of The Korean Institute of Communications Information Sciences.
☆ This paper has been handled by Prof. Hyukjoon Lee.

http://dx.doi.org/10.1016/j.icte.2016.10.003

The authors in [4] assume that users are differently served according to their own class. The authors in [5] present stochastic integer programming, and the authors in [6] propose a heuristic algorithm for the high complexity of integer programming.

This kind of approach using reserved instance derives the optimal configuration of reserved instances to minimize the total price, which

requires the expected usage of instances. Thus, the authors in [7] build a forecasting model to predict the usage of instances and conduct the optimization model.

Previous works have considered various functionalities; however, some of them are not suitable for use in actual environments. The prediction of the usage of instances is not well defined for the computing instances in cloud systems because actual usage fluctuates too much to predict. Instead, we restrict the usage of each reserved instance not to exceed a user-defined limit. This is actually very useful although it requires user intervention. We derive the optimization problem of heavy data by converting integer linear programming to linear programming. Ribon is a complete system that includes crawled data, data manipulation and optimization system, and dashboard for user interface. It is open source and available on GitHub to allow access and contribution by anyone.

3. System architecture

3.1. Overall system

Using the simulation tool in Fig. 1, we first collect price information from the AWS web sites. A total of 2118 EC2 instances were collected for evaluation. There are currently six pricing policies as follows.

- On-Demand: Pay at an hourly rate.
- No Upfront (1-year term): Pay at an hourly rate for 1-year RIs (Reserved Instances).
- Partial Upfront (1-year term): Pay for a low upfront and at an hourly rate for 1-year RIs.
- All Upfront (1-year term): Pay for 1-year RIs with one upfront payment.
- Partial Upfront (3-year term): Pay for a low upfront and at an hourly rate for 3-year RIs.
- All Upfront (3-year term): Pay for 3-year RIs with one upfront payment.

Then, the system obtains the expected usage of instances empirically and the previous used instances from a user, and validates their formats and the existence of price information. Based on the inputs from the user and the price information, the system obtains the cost-optimized configuration of instances and displays the results. The simulation tool has been developed using R, for data manipulation and optimization, and Shiny, for web dashboard.

3.2. RI optimization

This subsection described the formulation of the cost-optimized configuration and the solution. We define T as the

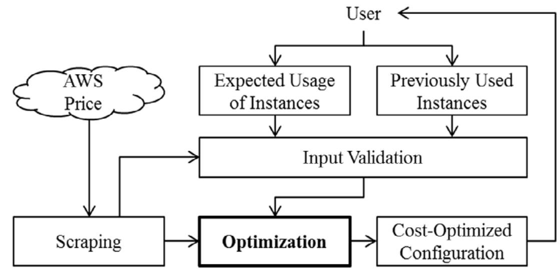

Fig. 1. Simulation tool system structure.

period of simulation; $N(t)$ as the expected usage of instances at month t; $m_\chi(t)$ as the number of instances χ to be used at month t; $n_\chi(t)$ as the number of instances χ to be contracted at month t; p_χ as the monthly price of instance χ; q_χ as the upfront price of instance χ, where χ is one of the six pricing policies $\{OD, 1N, 1P, 1A, 3P, 3A\}$; these are on-demand, 1-year no upfront, 1-year partial upfront, 1-year all upfront, 3-year partial upfront, and 3-year all upfront, respectively. Then, we can set the objective function, $f(m_{OD}(t), n_\chi(t))$, to be minimized as the total price at month t, which is given as follows.

$$
\begin{aligned}
f(m_{OD}(t), n_\chi(t)) = \sum_{t=1}^{T}\{ & m_{OD}(t)\cdot p_{OD} + m_{1N}(t)\cdot p_{1N} \\
& + m_{1P}(t)\cdot p_{1P} + m_{3P}(t)\cdot p_{3P} \\
& + n_{1A}(t)\cdot q_{1A} + n_{3A}(t)\cdot q_{3A} \\
& + n_{1P}(t)\cdot q_{1P} + n_{3P}(t)\cdot q_{3P}\},
\end{aligned}
$$

$$m_{1N}(t) = \sum_{\tau=t-11}^{t} n_{1N}(\tau),$$

$$m_{1P}(t) = \sum_{\tau=t-11}^{t} n_{1P}(\tau),$$

$$m_{3P}(t) = \sum_{\tau=t-35}^{t} n_{3P}(\tau).$$

One constraint is that the total number of instances to be used at month t is greater than or equal to the number of expected instances to be used, which can be expressed as follows.

$$m_{OD}(t) + m_{1N}(t) + m_{1P}(t) + m_{3P}(t) + \sum_{\tau=t-11}^{t} n_{1A}(\tau) + \sum_{\tau=t-35}^{t} n_{3A}(\tau) \geq N(t).$$

Moreover, the number of instances to be used and contracted should be greater than or equal to 0.

$$m_{OD}(t),\ n_\chi(t) \geq 0.$$

Additionally, we restrict the number of RIs such that it does not exceed a user-defined percentage r_χ as follows.

$$\sum_{t=1}^{T}\sum_{\tau=t-11}^{t} n_\chi(\tau) \leq floor\left(\sum_{t=1}^{T} N(t)\cdot r_\chi\right) \text{ (1 year term)}$$

$$\sum_{t=1}^{T}\sum_{\tau=t-35}^{t} n_\chi(\tau) \leq floor\left(\sum_{t=1}^{T} N(t)\cdot r_\chi\right) \text{ (3 years term)}$$