Future Generation Computer Systems 101 (2019) 502–523

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Value-based cloud price modeling for segmented business to business market

Caesar Wu *, Rajkumar Buyya, Kotagiri Ramamohanarao

Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Australia

HIGHLIGHTS

- Cloud pricing should consider both cloud service providers' cost and customers' value propositions.
- Cloud market segmentation can differentiate various customers' values.
- Value-based cloud pricing provides a better solution for a cloud service provider to maximize its profit.
- Four types of cloud pricing models illustrate a comprehensive framework of the cloud pricing process.
- Genetic Algorithm offers a convenient solution of optimizing each optimal price for each pricing model.

GRAPHICAL ABSTRACT

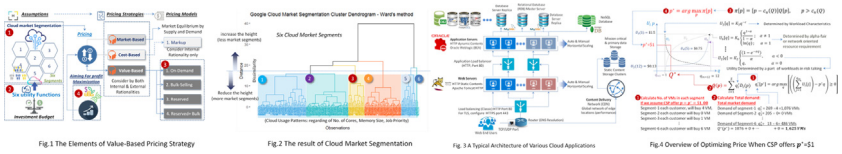

Fig.1 The Elements of Value-Based Pricing Strategy

Fig.2 The result of Cloud Market Segmentation

Fig. 3 A Typical Architecture of Various Cloud Applications

Fig.4 Overview of Optimizing Price When CSP offers p^*=$1

ARTICLE INFO

Article history:
Received 23 January 2019
Received in revised form 2 May 2019
Accepted 9 June 2019
Available online 26 June 2019

Keywords:
Customer utility functions
Cloud pricing models
Genetic algorithm
Price optimization
Cloud market segments
Cloud business profit

ABSTRACT

Cloud price modeling is the major challenge facing many cloud computing practitioners and researchers in the field of cloud economics, which is also known as "Cloudonomics." Previous attempts mainly focused on a uniform market and used existing price models to explain the issue of revenue maximization for cloud service providers (CSPs) from a cost or internal rationality perspective but paid less attention to the cloud market segmentation for cloud business customers from a surplus value or external rationality perspective. This study considers both aspects of the value proposition. Based on the assumptions of the customers' utility values for different market segments, we establish a framework of value-based pricing strategy and demystify the process of modeling and optimizing cloud prices for CSP to maximize its profits. This framework is built upon the theory of value co-creation for both customers and CSPs to form a business partnership. We show how to create four cloud pricing models, namely: on-demand, bulk-selling, reserved, and bulk + reserved. We also demonstrate how to identify the optimal price point of each model to maximize CSP's profit by genetic algorithm. We exhibit that reserved, bulk + reserved, on-demand, and bulk-selling can deliver a profit margin of 203%, 183%, 166%, and 157% for CSPs respectively. While the reserved model provides the highest profit margin, it does not necessarily mean that CSPs should adopt one model only. We provide a novel solution that allows CSPs to achieve the maximum profit by offering multiple pricing models simultaneously to various customers in the segmented market. We argue CSPs should capitalize on cloud pricing rather than price to gain more cloud market share and profit. Thus, we present state-of-the-art cloud pricing for segmented business to business cloud market.

1. Introduction

Value-based cloud price modeling for different cloud market segments [1] is vital to all Cloud Service Providers (CSPs) as it

* Corresponding author.
E-mail addresses: caesar.wu@computer.org (C. Wu), rbuyya@unimelb.edu.au (R. Buyya), kotagiri@unimelb.edu.au (K. Ramamohanarao).

https://doi.org/10.1016/j.future.2019.06.013
0167-739X/

will not only impact on CSP's profitability but also determine the sustainability of CSP's cloud business [2]. The goal of this study is to develop a comprehensive process framework of value-based price modeling that enables CSP to gain more cloud B2B market share for its profit maximization. Many previous studies can be considered as either cost-based or cost-plus models [3], which they were dependent on an assumption of a uniform market and paid less attention to the segmented market that carries heterogeneous values of customers. Furthermore, their processes of modeling mainly explained how to leverage two or three existing models (e.g., on-demand and spot instance) for CSP to maximize its revenue, which was subjective to a cloud capacity constraint that is equivalent to a cost. Subsequently, those works led to the issue of the internal rationality only.

The term of "Rational" means a decision is made according to reason or logic. In economics, people are assumed to be rational because they will systematically and purposefully do the best they can do achieve their purposes, given the available choices [4]. "Internal rationality" implies that a decision maker focuses on internal justification; for instance, a cloud price is determined by cost. In contrast, "external rationality" suggests that a decision should be made by an explanation of external factors, which a price is dependent on customer willingness to pay. In economics, it is essential that the pricing model is built upon the assumption that the individual is rational because people can be irrational.

The questions of how to create a cloud price model itself based on the business customers' value proposition and how to target the segmented market, especially, business to business (B2B) market have remained either unanswered or incomplete. To overcome this gap, we develop cloud price models that include both external and internal rationalities. For the external rationality, we examine two essential external factors, namely, cloud customers' utility values and B2B market segments. For the internal rationality, we take into account of CSP's cloud infrastructure cost. Based on our result of price modeling, we then use a genetic algorithm (GA) to identify the optimal price point of each price model for CSP to maximize its profit. One of the useful properties of GA is that it can solve a complex profit equation for intertwined variables without knowing the details of sub-functions. It is also convenient to upgrade the optimal price point of each model so that the process of price modeling can cope with the decision variation of cloud business strategy.

To demonstrate the process of value-based cloud price modeling, we exhibit and analyze different models, namely cost-plus, on-demand, bulk-selling, reserved (two-part tariff), and reserved + bulk for profit margin comparison. The cost-plus pricing models are often prevalent [2] because "they carry an aura of financial prudence… to yield a fair return on overall costs (or resources), fully and fairly allocated". However, these models fail to capture heterogeneous values of customers. In contrast, four value-based models can reflect the value proposition of both cloud customers and CSPs. Those models can be considered as "value co-creation" [5,6] because CSPs are seeking a partnership with their cloud customers in the cloud market value chain. We show these models allow CSPs not only to satisfy customers' needs but also to achieve a better profit margin in comparison with the cost-based model. Overall, we provide a process solution that has a quantitative measurement under a single currency (or business revenue contribution) can capture different cloud customer service metrics (e.g., increase sales, customer retention, investment efficiency, maintain a specified SLA, reduce checkout queueing time, etc.). To better illustrate the entire process of modeling, we use the following scenario to explain the details.

1.1. Background

Assume a group of decision-makers of a hosting firm decide to expand its hosting business into the cloud B2B market. It implies that the firm wants to become a new CSP to compete with other existing CSPs (either global or local CSPs). If the initial investment budget (both capital and operation expenditure or Capex and Opex) and business goals (targeted revenue, profit, and market) have been approximately identified, the decision makers want to know how to achieve the business goals. There are two fundamental questions must be clarified: "How does the firm form the right pricing strategy for the identified business goal?" and "how does it decide the appreciated cloud price models along with optimal price points, sales volumes, and unit cost to achieve the maximum profit?" These questions will help the CSP to divert its limited resources (investment budget and technology expertise) to serve its targeted customers better so that it can maintain its cloud business profitability and sustainability. There are many possible pricing strategies to reach the business goal, namely cost-based, market-based, and value-based pricing. As Hinterhuber [7] indicated, both cost-based (37%) and market-based pricing (44%) are much popular than value-based pricing (17%). Nagle [2] observed that historically, cost-based pricing is the most common pricing strategy in most industries because "in theory, it is a simple guide to profitability; in practice, it is a blueprint for mediocre financial performance". Unfortunately, the issue of the cost-based pricing strategy is when there is strong market demand, the average unit cost will decline, and the price reduction should follow because the profit margin is determined by the unit cost (e.g., 30%–100%). Conversely, when the market demand becomes weak, the average unit cost will go up, and the price should be raised. It contradicts a sensible pricing strategy in term of market response.

The alternative way of cost-based pricing is either market-based (or competition-based) or value-based (customer-driven) pricing. Market-based pricing is to set cloud service price based on the current competition condition or equilibrium of supply and demand. However, competition-based pricing could mislead CSPs to see market-based pricing as a zero-sum game, which what the customers' gain is the CSP's loss [8]. They might also believe they do not influence price because market-based pricing is a competition behavior of the market. In contrast, value-based pricing can offer customer needs and create real value to satisfy customers because it is determined by customers' utility values.

Nonetheless, the definition of value-based pricing can be subject to a wide range of interpretations. It is dependent on the context of the content. The term is often defined as a pricing process for an individual's preference (ordinal utility [9]) that aims to the B2C market. However, this paper of value-based pricing focuses on the marginal value (cardinal utility) that aims to the B2B market. It implies the process is to capture a proportional of value that CSP might impact on the targeted cloud customers for their business [8]. In other words, a CSP is to develop and deliver the cloud service values for its cloud customers to achieve business success and then seek a reward for its distributed services.

In general, the B2B market emphasizes the entire value chain and partnership development. The purchasing decision is not made by single or few individuals, but by more than dozens of stakeholders for the cloud service values that CSP can offer. Therefore, value-based pricing becomes one of the effective pricing strategies for the cloud B2B market. The cloud services can influence the customer's business in term of increasing their profit margin, higher business revenue and lower the operation cost.

Overall, the framework of value-based pricing strategy includes (1) **identifying** target customers and workload patterns

that are related to each cloud market segment, (2) **quantifying** cloud customer utility functions that are associated with service values and cloud service metrics, (3) **establishing** various cloud pricing models based on the specified customer's utility functions, (4) **identifying** the optimal price points for CSP to achieve the total maximization profit from all market segments. Fig. 1 presents this processing framework of price modeling of all elements.

Due to the limited space, it is impossible to include all four elements of a value-based pricing strategy into a single paper. This study will only focus on developing cloud pricing models (element 3) and identifying the optimal price points (element 4). The other elements of the value-based strategy have been presented in separated papers, and one of them has been published [10]. However, we will lay out enough details of cloud market segmentation (element 1) and customer utility functions (element 2) [11].

1.2. Cloud market segmentation

The purpose of cloud market segmentation is to gather cloud customers' usage patterns so that a CSP can work out a right pricing strategy to serve its targeted customers well while it can achieve its maximization profit within its budget constraints. In fact, Yankelovich [12] specified the detail criteria of market segmentation: (1) align with the company's strategy, (2) specify where the revenue and profit come from, (3) articulate customer's business values, (4) focus on actual business behaviors, (5) make sense to the firm's executive team and the board and (6) to be flexible to accommodate or anticipate market changes quickly. According to these market segmentation criteria, we develop a novel solution [10] that allows CSP to identify the cloud B2B market segment quickly. The solution is a combination of hierarchical clustering (HC) with time-series (TS) methods according to two datasets, which one is from Google public dataset [13] and other is extracted from a local hosting firm for its hosting business. From Google's dataset, we can develop six potential cloud market segments that are determined by the number of parameters of usage patterns, such as job priority, cores quantity, memory size, and AMD's virtualization workload guidelines [14]. This number of cloud market segments is within the range of McDonald's suggestion [15], which the suggested number of the market segment is between 5 and 10.

The results of cloud market segmentation are shown as in both Table 1 and Fig. 2. The details discussion of these market segments associated with utility function has been presented our previous publications [10,11]. Once the cloud market segments have been quantified, the next issue is how to develop the cloud customers' utility functions for these cloud market segments.

1.3. Modeling cloud customers utility functions

The goal of modeling cloud customers' utility function is to quantify the cloud customers' experiences and preferences (utility values) that are subject to the cloud resources provision. Practically, these subjective experiences concern the running applications for cloud customers to generate business revenue or profit, which can be quantified by the service metrics [19].

The meaning of utility is quite ambiguous because it consists of different connotations. Historically, the implication of utility was derived from utilitarianism. It means a subjective experience and satisfaction. It is known as the utilitarian tradition. Later, this term has been extended to the contractarian tradition, which emphasized social welfare [20]. As a result, the contemporary meaning of utility has three connotations:

(1) The economic utility refers to subjective satisfaction and happiness. "It is an alternative way to describe preference and optimization" [4] The utility value in this context is measured by different preferences under information uncertainty in term of risks and wealth.
(2) Another implication of utility is an essential infrastructure service for the public. Sometimes, it is also called as "public utility", such as water, electricity, and telephone service that are supported by some incumbent providers. It is associated with the term of social welfare.
(3) "Utility" also refers to the utilization rate. It is measured by a percentage value between 0 and 1. For example, the utility of a network means its utilization rate. It is a concept of efficiency. It is different from the economic connotation of utility that is measured by preferences.

However, there are many research works that assume both economic utility and utilization rate are the same. The utilization rate can be included in a cloud service metric, but it is not the same as the utility value in an economic sense. Economically, a business customer's utility represents the amount of business revenue or profit that is contributed by the number of VMs (e.g., wealth). For instance, the utility of a mission-critical application will be totally different with the utility of backend type of workloads, such as log data processing or MapReduce [19] because the end users will pay a different price for the cloud services. The question is, how we can use a single currency to reflect various utility values and align with CSP's profitability? To solve this issue is to unify all customers' utility values and CSP's profit into a measurement of cloud customers' business revenue or profit. This is also known as value co-creation. The benefits of value co-creation are that CSPs can reduce investment risk and maintain cloud customer loyalty [21] and uphold CSP's profitability and business sustainability. The modeling process of quantifying customers' utilities is to establish a relationship between the customer's business revenue or profit contribution (a dependent variable) and the number of VMs (independent variable) to be provisioned.

Based on different characteristics [22] of the cloud business applications, we organize utility functions into three categories:

- Utility functions (Segment 4 and 5) are defined by High Availability (HA) characteristics [23–25].
- Utility functions (Segment 1 and 3) are determined by response time characteristics [26].
- Utility functions (Segment 2 and 6) are identified by risk characteristics (risk-averse, risk-seeking, and risk-neutral [19]).

The process of how to quantify these utility functions is presented in the paper [11]. Table 2 highlights the result of six utility functions. (Details assumptions of these functions are presented in Section 3.2.1.) Now, the subsequent questions are how we can build various price models for a CSP to capture more cloud market share and how to identify the optimal price point of each model for profit maximization? These problems are what we will focus on in this research.

1.4. Problem definition and solution

By microeconomics [27], we can formalize the CSP's profit problem into the following equations. Eqs. (1) and (3) mean the total business profit is dependent on a sales price, an average unit cost (or a marginal cost), and sales quantity (e.g., market demand). Intricately, the quantity is a function of a price, and the

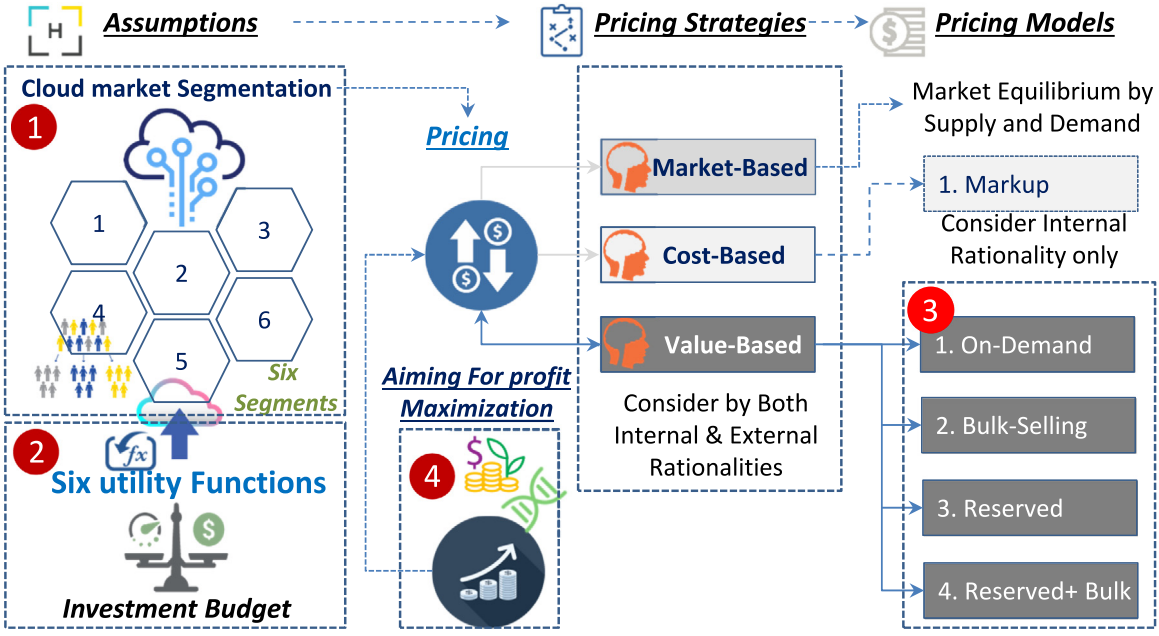

Fig. 1. The elements of value-based pricing strategy.

Table 1
Cloud customers utility functions and market segments.[1]

Segment	1	2	3	4	5	6	Total
Average job priority[a]	1	0	2	0	3	3	
Average number of cores	2	23	1	1	3	13	
Average number of memory	7	6	6	3	86	102	
Percentage	30.1%	23.0%	10.0%	26.3%	9.1%	1.4%	100%
Predicted sales vol[b]	269	205	90	235	81	13	893
Estimated possible workload[c]	Mainly Static	Backend[d]	Static & Dynamic	HA	HA	Dynamic Content	
Example of Apps	Virtualized Desktop Infrastructure, Email Server	MapReduce, log analysis File & Print	Web Hosting Server & Online checkout	Disaster Recovery	Database Backup & Terminal Server, SLA	Dynamic Content Delivery, Terminal Workload	

[a]In this case, "job priority" carriers more weight for the decision of cloud workload pattern [16].
[b]Sales Volume is estimated by time series (TS) predication without consideration of probability, which will be done in separated research work.
[c]The possible workload estimation is based on the recommendation of AMD's paper and cloud design patterns [17] [16] [18] [15].
[d]Backend type of workload patterns might also include business intelligent (BI) or log data analysis [16].

price is an inverse function of the quantity. Mathematically, we can present this interdependent relationship in Eq. (2)

$$\pi\,[p] = R\,[p] - C\,[Q\,(p)] \tag{1}$$

$$C\,[Q\,(p)] = c_u\,[Q\,(p)]\ *\ Q\,(p)\,,$$
$$R\,[p] = p * Q\,(p)\,, \qquad p = Q^{-1}\,(p) \tag{2}$$

$$\pi\,[p] = Q\,(p) * (p - c_u\,[Q\,(p)]) = pQ\,(p) - C\,[Q\,(p)] \tag{3}$$

where $\pi\,[p]$ is a cloud business profit, $R\,[p]$ is a cloud revenue, $C\,[Q\,(p)]$ is the total cost, p is a unit price and $c_u\,[Q\,(p)]$ is the average unit (or marginal cost) which is also a function of the total sales quantity $Q\,(p)$.

The issue is how we can achieve the maximum profit by identifying the optimal price point Eq. (4). While the equation appears evident and straightforward, it is difficult to find a clear solution because of both functions $Q(p)$ and $p = Q^{-1}\,(p)$ are generally unknown

$$p^* = \underset{p}{\operatorname{argmax}}\ \pi\,[p] \tag{4}$$

The primary challenge is that the relationship of $p = Q^{-1}(p)$, $c_u\,[Q\,(p)]$, and $Q\,(p)$ is intertwined. Moreover, these equations will become progressively more complex if various pricing models are introduced.

Previous works solve the problem by excluding the cost component from a profit Eq. (1) [28] or by making some restricted assumptions [29] [30], or by assuming a uniform market that is derived from α-fair utility [31]. Others assume a price is a simple linear equation based on the AWS' historical data within a coefficient band [32]. Still, others intend to offer a solution by mixing with existing on-demand, reserved and spot instances from a CSP's perspective [29]. Although their works have made

[1] HA = High Availability, DR= Disaster Recovery, VDI = Virtual Desktop Infrastructure

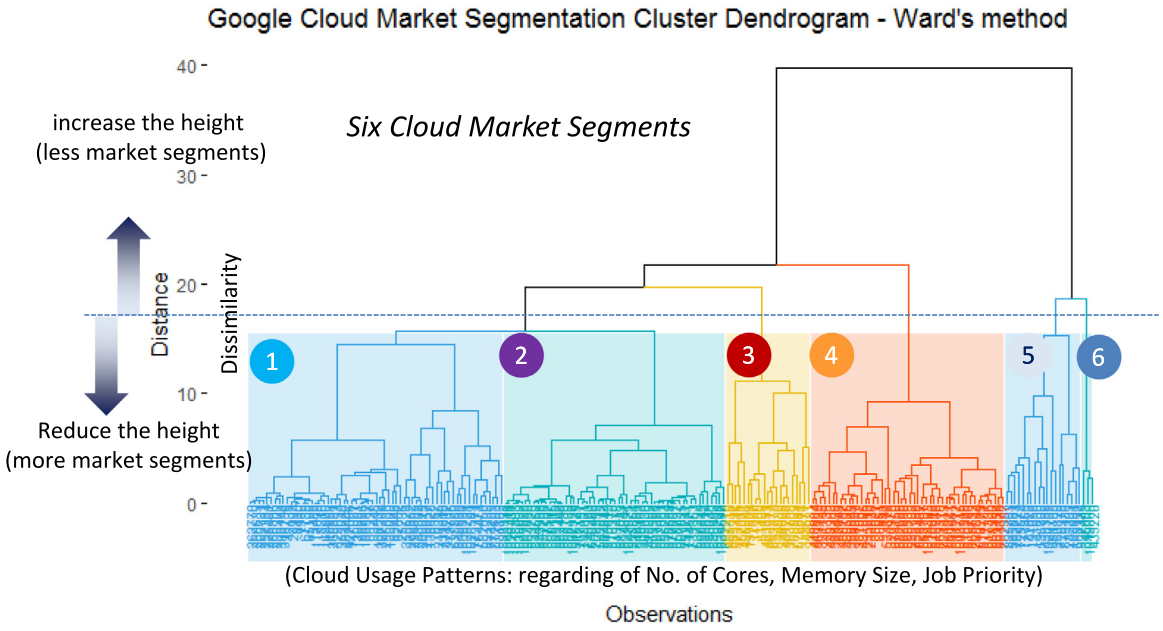

Fig. 2. The result of cloud market segmentation.

Table 2
Cloud customers utility functions and market segments.

Business application workloads	Market segment "i".	Cloud customers' utility Functions$U_i\,(q)$[a]
Virtual Desktop Infrastructure (VDI)	1	$U_1\,(q) = K_1\,(q_m + rq)\,, r < 0$
Backend Data Processing, Big Data	2	$U_2\,(q) = K_2 \begin{cases} \frac{(1-e^{-\alpha q})}{\alpha}, \alpha \neq 0, \alpha < 0 \\ q, \alpha = 0 \end{cases}$
Web Hosting or Online Checkout	3	$U_3\,(q) = K_3 q^{-c}$
Disaster Recovery or HA	4	$U_4\,(q) = \begin{cases} \theta K_5 & 1 \leq q \leq \mathrm{k} \\ 0 & k \leq q \leq q_{\mathrm{m}} \end{cases}$
SLA Backup, High Availability (HA)	5	$U_5\,(q) = \begin{cases} K_5, 1 \leq q \leq k \\ 0, k < q \leq q_m \end{cases}$
Content Delivery	6	$U_2\,(q) = K_6 \begin{cases} \frac{q^{1-\alpha}}{1-\alpha}, \alpha \neq 1 \\ \ln\,(q)\,, \alpha = 1 \end{cases}$

[a] $U_i(q)$ is a utility function in each cloud market segment. K is a scaling factor that is determined by cloud customers' business (explain later). α is the degree of risk preference for a performance, q is the quantity of Virtual Machine(VM), and q_m is the maximum number of VM that cloud customers will purchase. This is a arbitrary number. It can be 10 or 20. It is just the matter of a scale.

excellent progress in the context of cloud price modeling for the B2C market, many critical aspects of modeling remain unanswered. This study provides various solutions to resolve these issues. These solutions encapsulate the comprehensive process framework of value-based pricing strategy.

1.5. Contributions

By providing the various solutions, this work has made the number of contributions:

- To the best of our knowledge, it is the first time to create various cloud price models based on market segmentation theory and the number of utility functions that are defined by cloud customer business revenue or profit contribution.
- This work has clearly illustrated how to establish four value-based price models according to the defined business strategy
- By leveraging the actual retail pricing experiences, this study develops bulk-selling and reserved models for CSP to have more pricing options to achieve a higher profit margin.
- This work also illustrates the relationship between bulk-selling and bundle services. By developing various cloud pricing models, CSPs can spontaneously launch more pricing models to capture more profit across various market segments.
- We demonstrate how to apply GA to identify the optimal price point for each price model.
- The price models are dependent on both internal (CSP's cloud infrastructure costs) and external (cloud market segments and customer utilities) rationality.
- This paper presents novel and practical solutions so that many practitioners can plug in their datasets and build their own price models based on the defined company's business strategy.
- Most importantly, this study shows how to calculate the total revenue and profit based on different pricing models that are offered to various customers spontaneously.

1.6. Paper organization

The rest of the paper is organized as follows: Section 2 provides a brief overview of related works in cloud pricing. Section 3 formalizes four value-based pricing models according to various assumptions with different constraints. Section 4 presents concise information about genetic algorithms (GA) and how to determine the GA parameters for our experiments. Section 5 shows the experimental results. Section 6 offers a detailed analysis of cloud pricing and optimization. Section 7 concludes the paper and proposes future research directions.

2. Related work

In light of the value theory [33], we can approximately classify most of cloud price models into three basic categories, namely

value-based, market-based, and cost-based pricing. The value-based pricing is often considered as a subjective view of the cloud pricing from a demand side because it concerns the measurement of customers' subjective experience and utility preferences. The cost-based pricing is regarded as an objective view of cloud pricing from a supply side because it is built on the physical quantity of a unit cost. The market-based pricing is an interactive view of both subjective and objective for the equilibrium of supply and demand at the marketplace. According to this classification, we can probably classify most of the literature on the topic of cloud pricing models as either market-based or cost-based models [34].

For example, Macias et al. [35] used a genetic algorithm method determining a cloud price. Their model can be considered as market-based pricing. The study aims to offer a solution to a competitive price for a negotiation of the services market. However, they recognized that their work has some limitations. They believe "it is difficult to establish a profitable pricing function". We show how to overcome this limitation and bridge this gap in Section 3. Although Macias et al. [35] made some progress in term of modeling the cloud utility function for SLA metrics, one of the critical issues has remained unsolved, which is how to include the demand side's utilities for CSPs to generate various cloud pricing models and to achieve a partnership with cloud customers [36] in a cloud market value chain.

Kilcioglu et al. [37] present a calibrated benchmark model for cloud pricing based on empirical data. Their model can also be categorized as one of the market-based pricing models. Kilcioglu et al. [37] explained the market trends of the cloud price and higher profit margin of AWS based on the quality competition assumptions under both monopoly and duopoly market environment. The paper showed that the utility function of the cloud customer consists of three elements, subjective values, delay sensitivity, and service quality.

It was the first time that the demand side's utility function had been defined as a function of both subjective values and objective costs [37]. The paper made important contributions to the theoretical modeling of price-quality competition in both monopoly and duopoly competition market. Nevertheless, many problems are still unanswered, such as the determination of subjective values of utility for the cloud B2B market.

Aazam et al. [30] established a resource-based price model by cloud customer's historical pricing record for digital media stream workload across an inter-cloud environment (via cloud brokers). Although the authors made a great effort to build a model equation for inter-cloud pricing, many critical values of the equations are restricted to a particular case, e.g., a data stream type of workload. Nonetheless, the paper provided a framework of modeling and analyzing AWS on-demand and reserved instance pricing based on historical observation.

Yeo et al. [38] argued that automatic metered pricing model for a utility computing service (computing service as a commodity) could achieve a better revenue result in comparison with fixed pricing, fixed-time, and Libra [39] plus dollars $ [40] (a pricing model based on the users' requirements). The paper presented a compelling pricing model for self-justification, but, more experiments are required, as the authors indicated. Xu et al. [31] presented a similar idea and developed various pricing models (such as the 1st order discrimination, resource throttling, energy (or cost) saving and SLA charge) to maximize CSP's revenue that is subjective to CSPs cloud infrastructure capacity and customers' surplus value. The authors argued that the usage price depends on the utility level distribution and the elasticity parameters α on the base of their theoretical proof for Theorem 1 (see Eq. (5) by leveraging KKT condition [41]). Although their utility connotation was referred to as economic utility, the alpha (α) was derived from the α-fair network utility rather than a customer's preference. They concluded that pricing discrimination had no effect on CSP's revenue maximization.

$$p_v = \frac{\lambda}{1-\alpha} \tag{5}$$

This conclusion contradicts Claycamp and Massy's [1] the theory of market segmentation and McDonald's the practical solution of market segmentation [15]. There are some gaps in term of Xu's work.

(1) The economic sense of isoelastic utility function has different meanings of α-fair network utility because the former one is to measure a subjective experience and the latter one means the efficiency of utilization rate.
(2) The optimal price: p_v is dependent on the variable of Lagrange multiplier λ, which is not defined.
(3) As a result, the α-fair parameter is not inversely equal to price elasticity of demand of an isoelastic utility function.

$$E_d = \frac{\partial Q\,(\cdot)\,/Q\,(\cdot)}{\partial p/p} \tag{6}$$

where $Q\,(\cdot)$ is the quantity of the demand good. The α-fair utility means a priority of time scheduling while the α of isoelastic utility means the degree of risk. As Xu et al. [31] indicated their work was an extension of Hande et al. [42] study of pricing access networks with capacity constraints for revenue maximization.

Before Xu's paper, Joe-Wong and Sen [43] had also proposed a similar solution to a cloud pricing strategy that is subjective to the cloud capacity. The root of their pricing strategy was also derived from the access networks. The purpose of their research work was to develop an analytic framework to balance the fairness (welfare concept) of resource priority and CSP's revenue maximization by various pricing models. Although there were some differences between them, (e.g., Xu's work included a probability of utility level distribution, and Joe-Wong discussed fairness) both studies assumed there was a uniform market and corresponded to a α-fair utility function. All studies relied on the Lagrange multiplier or Karush–Kuhn–Tucker (KKT) conditions to identify the optimal price point, which is subjective to the specified limited capacity. Ultimately, they used the analytic tool to prove there is an optimal price point.

Recently, Shahrad et al. [44] proposed an incentive pricing solution by balancing limited cloud capacity and demand peak time. Shahrad's core idea is to leverage the cloud price as an incentive to regulate the usage behavior of cloud business customers, which means they would allocate cloud resources by themselves according to CSP's price variation. It is a self-regulate idea to eliminate its own demand during a peak time and fill its workloads during a valley time. The customers' utility function is the same as α-fair one.

All these studies assumed one type of utility function that is α-fair network utility for cloud customers. All papers assumed that economic utility and the utilization rate of a network are equivalent. In order to achieve maximum profit, the objective function has to be differentiable. In contrast to the α-fair network utility function, Chen et al. [45] proposed a utility function that is driven by the cloud customer's satisfaction in term of price and response time shown as follows:

$$U\,(p,t) = U_0 - \alpha p - \beta t \tag{7}$$

where U_0 is the maximum utility value and both α and β and constant coefficients. Price p and response time t are two independent variables to reflect different levels of utility value or customer satisfaction. If both p and t is equal to zero, it means the customer has maximum utility value. This is a linear utility function. Actually, the response time can be represented in price (or

a cost) p because if CSP provision more cloud resources, e.g., VMs for workload process, the response time t can be reduced. In addition to this issue, the paper did not give the optimal price point between CSP's profit margin and cloud customers' surplus value (customer satisfaction).

In comparison with building a pricing model from scratch, the works [28,29,32,46] focused on current cloud pricing models offered by different CSPs for profit maximization. Xu et al. [29] combined both reserved and spot instance prices that allow a CSP to maximize its revenue and profit through a dynamic cloud pricing model. The work was derived from empirical observation of the historical price of Amazon Web Services (AWS). The paper made contributions to an alternative pricing model for a spot pricing scheme. Following the similar line of reasoning, Alzhouri F. and Agarwal A. [46] constructed a theoretical or dynamic pricing scheme for CSPs to maximize their revenue via a solution of dynamic programming approach. The potential issue of their revenue maximization without consideration of average unit cost or marginal cost would become economically unsustainable. Toosi et al. [28] included all three types of pricing models, namely on-demand, upfront reserved and a spot for CSP's profit maximization but the unit cost of cloud resource remains untouched [47]. Brynjolfsson et al. [48] argued that this kind of cloud pricing could be "overly simplistic … blinding us to the real opportunities and challenges of cloud computing".

On the other hand, Agmon Ben-Yehuda et al. [32] suggested the price of AWS spot instance is not driven by some market mechanism or an auction approach rather than it is randomly generated from a close price range that has a dynamic hidden reserved price mechanism. This indicates that the price mechanism of AWS spot instance (2 min notification for termination) is similar to Google's preemptible VM instance (80% discount but terminated after 24 h execution time with 30 s notification), and Azure low-priority VM or eviction instance (with 60% (for Windows)-80% (other OS) discount, excluding B-Series VMs, 30 s notification). The problem with these instances (or VMs) is that both service availability and capacity cannot be guaranteed. Moreover, many new service features are excluded. Perhaps, MOZ's [49] business experiences[2] on 26-Sep-2011 provided a good lesson for many cloud business customers. The incident suggests that the spot instance is not designed for a mission-critical cloud application. Overall, we can summarize the main contributions, advantages, and potential gaps in these works in Table 3.

As Kash, I A. and Key P. B. suggested [50], the spot instance price model has been attracted much attention in the academic world for cost saving. Despite that, "the right answer remains unclear" [50]. One of the reasons is that many price schemes are restricted to a particular case or application. For example, Jain et al. [51] suggested a value-based price model by leveraging the spot instance discount, but the model is only designed for batch workloads. In other words, different models could have different purposes with different functions. To visualize all pricing models with different purposes and functions, we can use Table 4 to highlight their differences.

Although many researchers in this field have made excellent contributions to cloud economics, there are still many questions remain unanswered: such as **How to** generate more price models for various cloud applications that can capture more cloud market share? **How to** practically identify the optimal price point for each model? **How to** translate multiple dimension [50] of cloud service metrics (utility values) into a single currency between cloud customers and CSPs? **How to** address CSP's concerns of cloud B2B market? **How to** create a value co-creation solution for both cloud customers and CSP? **How to** determine the maximum profit with multiple pricing models? This paper, together with other our previous publications [10,11], provides a processing framework of a total solution for these questions.

[2] MOZ reserved bid for AWS spot instance was $2/per instance for more than 3 years

3. Cloud price modeling and models assumptions

3.1. Market assumptions

According to the theory of the B2B market [8], the cloud B2B market is a relational business market because it emphasizes building a mutual value-creation relationship or partnership with business customers. It requires long-term relationship development. In contrast, business to consumer (B2C) market mainly is focusing on the final transaction between a firm and an end-user [8]. From this perspective, we will first consider the cloud price models based on the assumption of a monopoly market [27] because the B2B market is much challenging for other market competitors to access to the existing market [52]. Furthermore, many innovative characteristics of cloud services often do have the existing market (However, the hedonic model [53] provides a possible solution to establish a cloud price model for the innovative cloud service characteristics). This premise is not prohibitive.

In addition to the monopoly assumption, we also assume the cloud market is not a uniform market rather than the segmented market because cloud customers have heterogeneous cloud applications. This assumption allows CSP to capture more cloud market share. Cloud market segmentation is to group personalized prices for heterogeneous demands so that the CSP can achieve the best profit margin within its resource capacity [1]. One of the typical examples of market segmentation in the service industry is the airline ticket price. The airline companies often classify their market into three or four segments, which is the 1st class, business class, economy, and cheap flights with different airfare prices and service conditions. Similarly, the cloud market can be grouped into different segments based on the different characteristics of cloud services.

3.2. Assumptions of quantifying VM resources

Following the segmentation result and the virtual server workload guidelines [14,18], we can approximately estimate the workload pattern of each cloud market segment, such as web hosting, high availability (HA), backend data process, disaster recovery, content delivery, etc. [54] as shown in Fig. 3. The VM quantity for one type of VM is represented by q, such as Amazon Web Service's instance of m4 extra-large or Google's ni-highmem-16. This quantity may vary from customer to customer. It is dependent on a type of VM instance and cloud business application. By consideration of all these factors, we set this maximum number is equal to 12 ($q_m = 12$) because we mainly focus on the small and medium enterprise (SME) customers so that this maximum quantity is justifiable for a typical SME's application. This number can be either increased or reduced. It is just a matter of a scale.

3.2.1. Pricing models assumptions

3.2.1.1. Cost assumptions. Along with the cloud market assumptions, we also assume the initial investment budget or Capital expenditure (Capex) for one type of VM. The Capex and Operational expenditure (Opex) ratio are 1:4. This ratio is based on local empirical data. The Capex is estimated by the latest average price of server hardware that is offered by major vendors, such as HP (HP Enterprise DL380, 2RU), Dell (PowerEdge R730), IBM (8203-E4A5634), and Cisco server (UCS M5). We also include some cloud data center installation costs [17], which are shown in Table 5.

Table 3
Summary of some previous works.

Category of pricing models	Main contribution	Advantages	Potential gaps
Toosi et al.'s Heuristic Algorithm of pricing model [28] (2014)	It combined three different pricing models for profit maximization for CSP profit maximization	Consider all available pricing models at that time	Excluded cost component.
Agmon Ben-Yehuda et al. Statistical regression (2013)	It provided a rough estimation of the AWS pricing model for spot instance	Proposed alternative solution for the pricing model	Observation of historical records. Lack of rationality
Hande et al. [42] α-fair utility model (2010)	It introduced one of the utility functions for the pricing model	Highlight price elasticity and utility function	Ambiguity definition of Utility and pricing Elasticity
Xu, Hong, and Baochun Li [31] α-fair utility model (2013)	It introduced the probability density function for cloud customer demands	Show KKT Proof	Contradict to Market segmentation theory
Joe-Wong et al. α-fair utility model [43] (2012)	It offered a mathematics framework for cloud pricing	Introduced multiple pricing models for cloud pricing	The only proof of the optimal price without consideration of market
Shahrad et al. [44], Cobb–Douglas $\rightarrow$ α-fair utility model (2017)	It proposed a novel idea of increasing cloud data center capacity utilization rate while to maximize CSP' profit	Show Euler homogeneous proof	Utility function has to be differentiable
Chen et al. Customers' Satisfaction linear Utility model [45] (2011)	It introduced a linear utility function for cloud pricing	It included both price and SLA level into the utility function	Not clear in term of the optimal solution for CSP's profit maximization

Table 4
Different pricing models comparison.

Purposes with various functions of Pricing model Comparison	Model explain	Model creation	Differentiable object function	Non-Differentiable	Market segmentation	Max. Rev.	Including cost element	Profit optimization	Optimizing algorithm	Optimal price point
Macias et al. [35]	√			√		√			√	√
Kilcoglu et al. [37]	√	√	√	√			√	√	√	√
Aazam et al. [30]	√							√		
Yeo CS. et al. [38]		√		√		√	√	√	√	
Xu et al. [31]		√	√			√			√	
Hande et al. [42]	√	√	√			√			√	√
Joe-Wong et al. [43]	√	√	√			√	√	√	√	
Shahrad et al. [44]	√	√	√			√	√	√	√	√
Chen et al. [45]	√	√	√			√	√	√		
Toosi et al. [28]	√					√			√	
Xu et al. [29]										
Alzhouri et al. [46]	√		√			√			√	√
Ben-Yeuda et al. [32]	√			√			√		√	
Kash et al. [50]	√	√							√	
Jain N [51]		√		√					√	
This model	√	√	√	√	√	√	√	√	√	√

Table 5
Cloud infrastructure cost assumptions.

Capex/per hour	Opex /per hour	Capex & Opex ratio	Number of physical servers	Configuration	Number of VMs capacity
$325	$1,300	1:4	400–600	8 or16 cores/per server	9,000–12,000

Note:
- Assumptions of investment Budget or Capex C = $3 million.
- The number of physical servers ≈400–600.
- The configuration per server is either 8–16 cores/ per server.

3.2.1.2. Utility function assumptions. From Table 1, we know the cloud market segment and predicted sales quantity, but we do not know the cloud customer utility function of each market segment. To optimize the cloud pricing models, we also need to define the cloud customer utility function for each cloud market segment. According to Krugman and Wells [55], the different individual would have different utility functions because different people would have different tastes and preferences. The essence of a utility function is to describe how people consume various quantities of goods in term of their subjective experiences and tastes by a less or more rational way.

If we assume CSPs just target the SME customers and focus on building mutual value generation; the modeling process is to define how their cloud resource (VM) can create SME's business profits. The effective modeling is that CSP should translate various cloud service metrics (Response time, SLA, end users retention, and leverage investment) into a single currency (business profit), which is also in line with CSP's business value proposition. As a

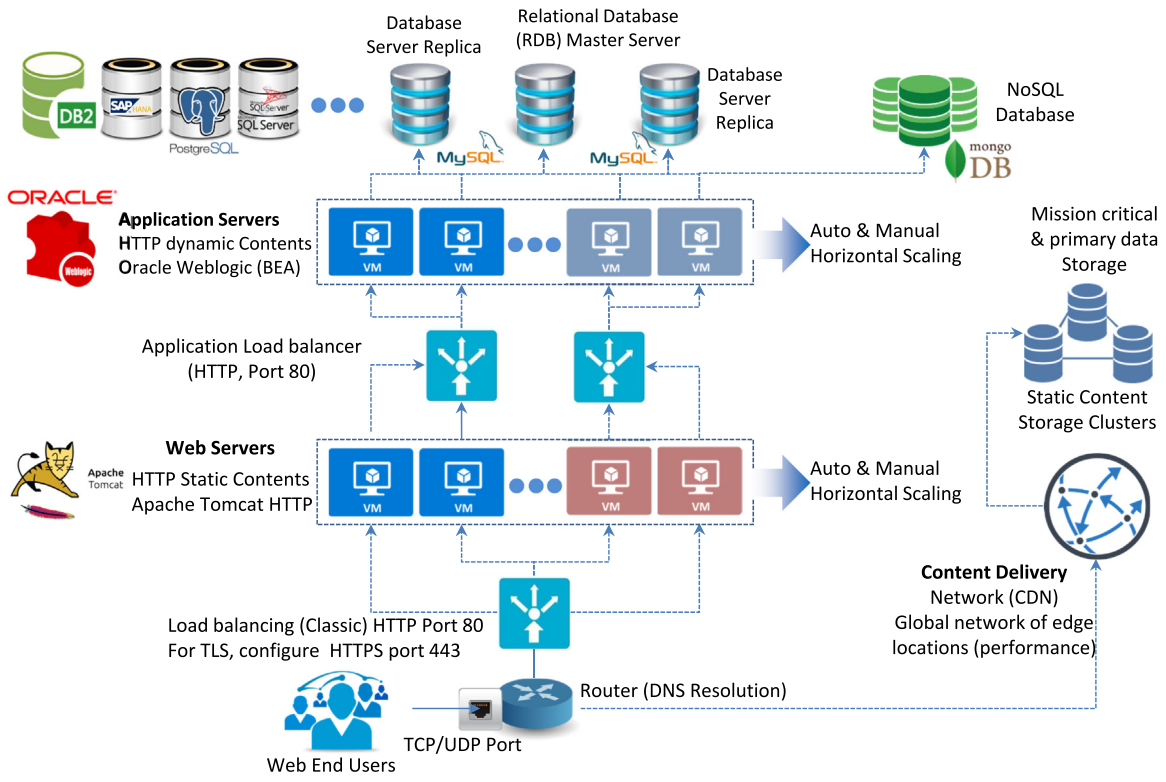

Fig. 3. A typical architecture of various cloud applications.

result, the cloud customer utility function is defined by the business customers' profit gain (surplus value) for cloud resources (or VM quantity) provisioning. We can use the following equations to describe their relation:

$$B_i = K_i \left(\sum_{q=1}^{q_m} u_i [q] \right), \quad i = 1 \cdots S \tag{8}$$

$$K_i = B_i / \left(\sum_{q=1}^{q_m} u_i [q] \right), \quad i = 1 \cdots S \tag{9}$$

where B_i is a yearly data. It represents customer business profit. If we check the Australian Bureau of Statistics (ABS) data [56] for small business, we can select a specific profit range for the targeted SME. If we just focus on the average net profit is approximately between $41,000 and $100,000 SME market, we can identify the values of B_i across all segments (Table 6). $U_i[q]$ is a customer's utility function for "i" market segment. "q" is the quantity that the customer will provision. K_i is the scaling coefficient that reflects the utility level that is associated with a cloud customer's business profit. (Further details will be illustrated in Fig. 4.)

3.2.1.3. Risk assessments. Risk assessments refer to a utility function is defined by cloud customers' preference for different levels of satisfaction for their business profit gain in term of their attitude towards risk of provisioning the various amount of VM resource. For example, according to the cloud customers' usage pattern, we can understand that the 6th market segment is for the cloud customers to deploy the web contents. It is a network-oriented utility function. According to [4,20,43,55,57], the iso-elastic utility function can describe the customers' utility in term of the cloud resources requirement (Eqs. (10) and (11)):

$$U_6 [q] = K_6 u_6 [q], \qquad u_6 [q] = \frac{q^{1-\alpha}}{1-\alpha} \; \alpha \neq 1 \tag{10}$$

$$K_6 = B_6 / \left(\sum_{q=1}^{q_m} u_i [q] \right) = B_6 / \left(\sum_{q=1}^{q_m} \frac{q^{1-\alpha}}{1-\alpha} \right) \tag{11}$$

where "q" is the number of VMs, and α is the constant coefficient. The coefficient α is also to measure the degree of relative risk aversion. In this case, we assume that cloud customers' utility value is dependent on the measurement of constant relative risk aversion (CRRA) [57] for content delivery applications workload. Based on a similar line of reasoning, we can also create the customers' utility function of the 2nd market segment as an exponential function [9].

$$U_2 (q) = K_2 \frac{\left(1 - e^{-\alpha q}\right)}{\alpha}, \qquad \alpha \neq 0 \tag{12}$$

We assume that the customers of this segment become risk-taking because the application (e.g., MapReduce) workload can be interrupted. Reliability and capacity guarantee of a cloud resource is not a significant issue. Cost saving becomes the main priority. Therefore, the coefficient value α is negative.

3.2.1.4. High availability. The high availability (HA) business applications require the mission-critical cloud infrastructure. If we assume the downtime should be less than 5 min/per annum, then the service level agreement (SLA) must be higher than five-9s (or 99.999%). Based on Markov Chain analysis [58], we can quantify the number of VMs required to guarantee SLA delivery. If the VM quantity is more than this specified number, the utility value will be diminished to zero. Moreover, all VMs have the same utility value because these VMs bundled together can guarantee SLA delivery. Consequently, we can define the utility function for the 5th segment as follows:

$$U_5 (q) = \begin{cases} K_5, & 1 \leq q \leq k \\ 0, & k < q \leq q_m \end{cases} \tag{13}$$

where k is the specified quantity of VM to guarantee cloud applications' SLA. q_m is the largest quantity that customers will

Table 6
Cloud customer surplus values (Profit contribution) in six market segments when $\boldsymbol{p}^* = \$1$.

Customer's profit or surplus $\boldsymbol{B_i}$	\$79,000	\$43,000	\$41,000	\$79,000	\$79,000	\$100,000	Total
Utility Functions $\boldsymbol{U_i(q)}$	$\boldsymbol{U}_1(\boldsymbol{q})$	$\boldsymbol{U}_2(\boldsymbol{q})$	$U_3(q)$	$U_4(q)$	$U_5(q)$	$U_6(q)$	
$q = 1$	\$1.50	\$0.01	\$1.50	\$0.75	\$1.50	\$0.29	
$q = 2$	\$1.36	\$0.03	\$0.75	\$0.75	\$1.50	\$0.45	
$q = 3$	\$1.23	\$0.05	\$0.50	\$0.75	\$1.50	\$0.60	
$q = 4$	\$1.09	\$0.12	\$0.38	\$0.75	\$1.50	\$0.72	
$q = 5$	\$0.95	\$0.18	\$0.30	\$0.75	\$1.50	\$0.84	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
$q = 11$	\$0.14	\$1.0.7	\$0.14	\$0.75	\$0.00	\$1.42	
$q_m = 12$	\$0.00	\$1.50	\$0.13	\$0.75	\$0.00	\$1.50	
Customers market demand[a] D_i	269	205	90	235	81	13	893
Cloud workload patterns	Virtualized Desktop Infrastructure, Email Server	MapReduce, log analysis File & Print	Web Hosting Server & Online checkout	Disaster Recovery	Database Backup & Terminal Server, SLA	Dynamic Content Delivery, Terminal Workload	

[a]The cloud market demand is derived from a hosting dataset based on time series [52]. These are addressable market.

provision [59]. Similarly, we can also build a similar utility function for the 4th market segment. The difference between 4th and 5th segments is the customers of the 4th segment might have its own existing cloud infrastructure. They will only provision a certain amount of cloud capacity if the price is below a specified threshold level θ in comparison with their own infrastructure costs.

$$U_4(q) = \theta K_4, \qquad 1 < q \leq q_m \tag{14}$$

3.2.1.5. Queueing time. In addition to the mission-critical workload applications, the utility function for the e-Commerce can also be modeled by a Markov Chain process. The basic idea of modeling the utility function for the 3rd segment is to reduce queueing time [60–62]. The following equation can define this type of utility function.

$$U_3(q) = K_3 q^{-c}, \tag{15}$$

where c is a constant value, we set the "c" is equal to 1 in this case because of the workload pattern (e.g., purchasing checkout). Alternatively, we can also use a linear function as a solution to describe the customer utility function for the 1st market segment of virtual desktop infrastructure (VDI). There are many VDI performance metrics of a hosting environment regarding users' experiences, such as the peak of Input/Output Per Second (IOPS), storage capacity, response time, Read/Write ratio, future growth, etc. If we assume these metrics [63] have been prefixed during the Proof of Concept (PoC) period before VDI rollout, the additional VM will only add a marginal cost to the cloud customers. So, we can use a linear model [45,64,65] to model the cloud customers' utility values.

$$U_1[q] = K_1(rq + q_m), \qquad r < 0 \tag{16}$$

where "r" is a constant, but it is negative, which reflects the diminishing return due to marginal cost increases.

3.3. Finding optimal price point for profit maximization

Once we have defined all utility functions for all market segments, we can create different price models for CSPs to maximize its profits. From an example in Fig. 4, we can illustrate a process of identifying the optimal price point of price model for CSP to maximize its profit.

Suppose a CSP offers \$1/per VM as its optimal price point (this price point is randomly selected. This one dollar could be the optimal price point for the CSP's profit maximization, but we do not know yet at this stage), we can calculate the cloud customer surplus values and a quantity of VM sales in each segment and the total market demand. According to the defined utility function of the 1st segment, the cloud customers will provision 4 VM, but not 5 VMs because 5 VMs would cost \$5 and the net surplus utility value of 5 VMs is only \$1.138, which is less than \$1.183 for 4VM. In other words, if each customer of the 1st market segment buys 4 VMs and the total number of cloud demand is 269, then the total sales volume of VM is 1076. Likewise, the customer of the 2nd market will not purchase any VM, but the 3rd segment will provision 1 VM. If we sum up all the VMs of all market segments, we can find the total volume of VM sales Q. As a result, we can calculate all the variables, including unit cost, profit margin, and total sales revenue. However, if this price point is not optimal, how can we find the optimal price point for the CSP to maximize its profit across all market segments? Before answering this question, let us think about "are there different pricing models to achieve a better profit margin?" This question takes us to the topic of building various cloud pricing models in comparison with cost-based pricing.

3.4. Markup pricing model

As Hinterhuber indicated [7], the cost-based pricing is still prevalent in most industries, which is over 37% of firms adopt it. If we assume the markup price is 100% of the average unit cost, the expected profit margin would be 100% Eq. (17). The process of determining a price is very straightforward. On the flip side, this pricing model could be either overshot or undershot due to the pricing without external rationality [27,66].

$$p[Q] = mc[Q(\cdot)] + \frac{Q(\cdot)}{|\partial Q(\cdot)/\partial p|} \tag{17}$$

where $p[Q(\cdot)]$ is the price, $mc[Q(\cdot)]$ is the marginal cost, $Q(\cdot)$ is the total demand quantity, $Q(\cdot)/|\partial Q(\cdot)/\partial p|$ is the markup price or profit margin. The price point is arbitrarily determined by the internal rationality or cost and a CSP's desired profit margin: $Q(\cdot)/|\partial Q(\cdot)/\partial p| = 100\%$.

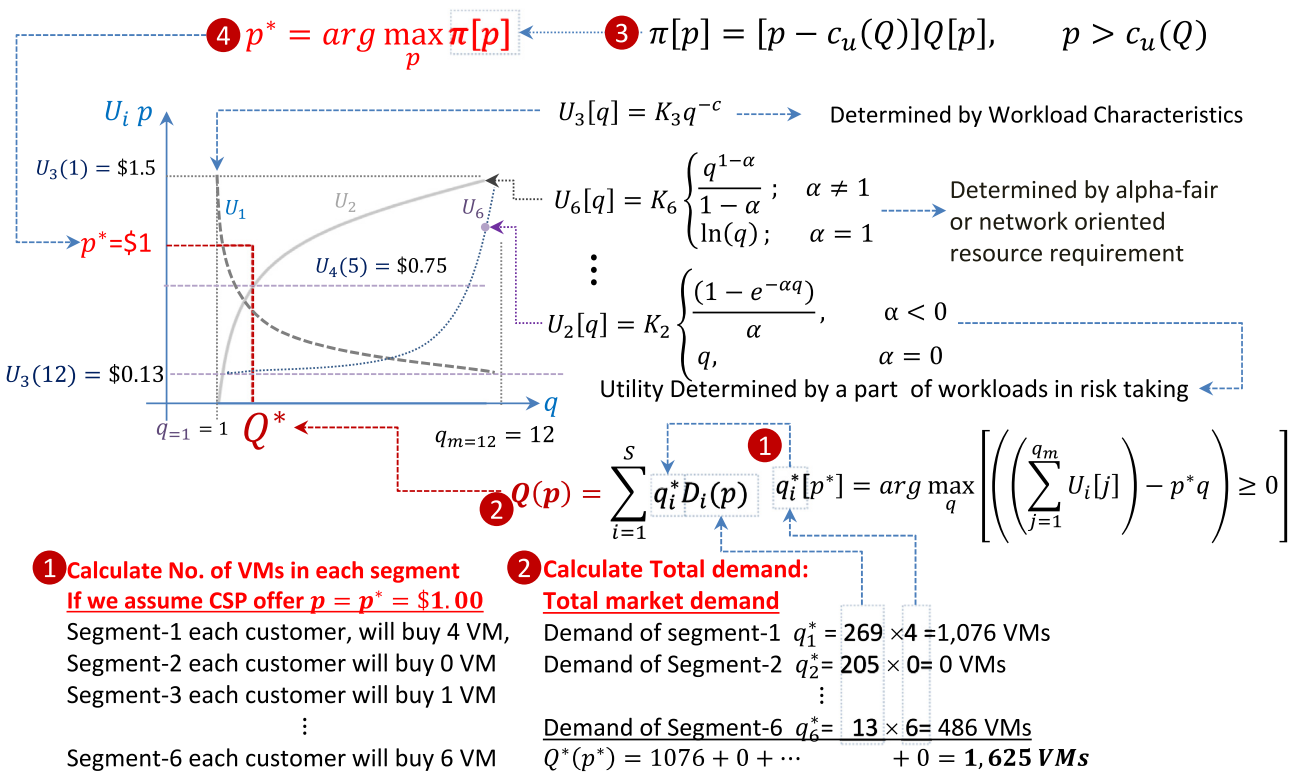

Fig. 4. Overview of optimizing price when CSP offers p^*=$1.

3.5. On-demand pricing model

Alternatively, we can create an "on-demand" price model [67] that is determined by both external and internal rationalities. Many leading CSPs offer this price model. It is also known as Pay as You Go (PAYG). Usually, CSPs would charge at an hourly unit-based price. While both Google Cloud Platform (GCP) and Microsoft Azure use a sub-hour rate. Azure is 1/60th hour or per minute base, and GCP is a 1/6th hour or per 10 min base [68]. The sub-hour price should give cloud customers more flexibility and scalability to run various types of cloud workloads for "on-demand". Our model adopts the hourly base unit. From the example of both Fig. 4 and utility functions are shown in Table 4, the following Eq. (18) can be generated to calculate the cloud customer surplus values (external rationality).

$$q_i[p] = arg\max_q \left[\left(\sum_{j=1}^{q_m} U_i[j]\right) - pq \geq 0\right] \qquad (18)$$

$$Q(p) = \sum_{i=1}^{S} q_i[p] D_i[p], \quad i = 1, \cdots, S$$

where S is the number of market segments, which is equal to 6 in this case. The q_i is the number of VM to be provisioned by the customers in the market segment "i". This quantity is decided by the customers' the maximum surplus value that is greater than zero for a given price p, which is offered by a CSP. q_i is a function of a price$p$.

$$\pi[p] = pQ(p) - C[Q(p)], \quad c_u[Q(p)] \leq p \leq M, \qquad (19)$$
$$c_u[Q(p)]Q(p) = C[Q(p)],$$

$$p^* = arg\max_p \pi[p] \qquad (20)$$

where $Q(p)$ is the summation of $q_i[p]$ of VMs multiplied by the estimated market demand $D_i[p]$ of each market segment. M is the normalized maximum utility value in Table 4. We generalize this value across all the segments ($1.5). $C[Q(p)]$ is the total cost based on the cost assumption of Table 3 (internal rationality). In summary, Eq. (18) is to determine the quantity $q_i[p]$ of VM in each market segment when customer surplus value is maximum. $Q(p)$ is the total VM sales for all market segments. Eq. (19) is the same as Eq. (3) in Section 1, with some specified unit cost (Refer to Table 5). Eq. (20) is to identify the optimal price for the profit maximization, which is the same as Eq. (4).

According to customer surplus values, some customers will provision a certain number of VMs, and others might not buy any for a given price per VM. It is dependent on the type of utility function $U_i[j]$ or customers' utility (external rationality) and CSP's offering price p and $c_u[Q(p)]$ per VM (internal rationality), which has been illustrated in Section 3.3. The question is "would it be possible to generate different type of pricing model so that the customers of both the 1st and the 2nd market segments will make a purchasing decision?" For example, if a CSP can offer a particular percentage discount on VM price but customers have to purchase VMs in a bulk size? This question leads to creating the bulk-selling pricing model.

3.6. Bulk-selling and service bundle pricing model

In comparison with on-demand, CSP can generate a bulk-selling or services-bundle pricing model. The goal of the bulk-selling is to encourage cloud customers to buy more for a better pricing deal. There are many examples of the bulk-selling price model, such as one of the retail giants, Costco Wholesale. The telco industry often uses service-bundle for different market segments. Service-bundling means bundling different types of services into one package and bulk-selling is to group different sizes of the same product or service into one package. Two models are closely related.

For example, one large and one extra-large size instances can be formed as one package, which is equivalent to 12 small VMs

for bulk-selling (see Fig. 5). According to [37] observations, the AWS price of the current size of the VM is equal to 2 power of "k" minus 1 and multiply by the price of the smallest or baseline VM size (where $k = 1, 2, \cdots$, the current size of VM). This pricing mechanism can be written as $p_k = 2^{k-1}p_0$ and p_0 is the price of the smallest VM size, p_k is the current size of VM. Such prices scheme is adopted by many CSPs for their majority types of VMs. The distinct advantage of adopting this pricing scheme is that the CSP can increase capacity flexibility by building a large VM resource pool with finer granularity and minimize a footprint of cloud infrastructure in a cloud data center. This means reducing the investment budget, increasing sales, and meeting the fluctuation demand for cloud resources.

Both bulk and bundle type of pricing scheme can be tailed for a particular business application, such as mission-critical workload, virtual data center, and Disaster Recovery, which the CSP will only sell for a fixed number of VMs in bulk. The cloud customers will decide whether to buy or not, based on their maximum utility (surplus) values. This bulk-selling model is built from equations from 21 to 26. If we assume the bulk-selling size to be "b", we can use "mod" function "B" to test whether any requested quantity matches the bulk-selling size or not. If it does not, then we can artificially set the customer surplus value to a negative value (for example, -200), which is to reject the customer's purchasing request Eq. (21). Otherwise, the customer's surplus value will be calculated Eq. (22), but it should be greater than zero Eq. (23).

$$\begin{aligned} &IF\ B = q - b\left\lfloor \tfrac{q}{b} \right\rfloor > 0 \rightarrow CS_i = -200, \\ &\forall b \in q = \{1, 2, \cdots, q_m\} \end{aligned} \tag{21}$$

Otherwise,

$$CS_i\left[p, q\left(b\right)\right] = \left(\left(\sum_{j=1}^{q} U_i\left[j\right]\right) - pq\left(b\right)\right) \geq 0, \tag{22}$$

$$q\left(b\right) = nb, \qquad n = 1, 2, \cdots$$

Then, comparing all surplus values in the market segment i and find the maximum value. Based on this maximum surplus value, the VM quantity q_i can be identified in the market segment i

$$q_i\left[p, b\right] = \underset{q}{\operatorname{argmax}}\ CS_i\left[p, q(b)\right] \tag{23}$$

Multiple market demand $D_i$with $q_i$in the market segment i Eq. (24) and sum up all quantities of market segments, then we can optimize both price p and b to find the maximize profit value (Eqs. (25) and (26))

$$Q\left(p, b\right) = \sum_{i=1}^{S} q_i\left[p, b\right] D_i\left[p\right] \tag{24}$$

$$\begin{aligned} &\pi\left[p, b\right] = pQ\left(p, b\right) - C\left[Q\left(p\right)\right], \quad c_u\left[Q\left(p, b\right)\right] \leq p \leq M, \\ &C\left[Q\left(p\right)\right] = c_u\left[Q\left(p, b\right)\right] Q\left(p, b\right) \end{aligned} \tag{25}$$

$$\left[p^*, b^*\right] = \underset{p,\ b}{\operatorname{argmax}}\ \pi\left[p, b\right] \tag{26}$$

For example, if a CSP offer the bulk size b is 4 and the VM price is \$0.5, the surplus values are set to negative ($CS_i = -200$) for all quantities of q that is not divisible by a package size ($b = 4$). Otherwise, the customer surplus value will be calculated. The maximum surplus value of the 1st market segment is 2.584, which is corresponding to the VM quantity of 4. There are other VM quantities (8, and 12) can be divisible by the package size, but the surplus value is either 2.184 or 0.00246. In comparison, purchasing 4 VMs has the maximum surplus values for cloud customers. Base on the same principle, the total sales volume for all segments can be found, which is equal to $Q\left(p, b\right) = 6136$ Eq. (24). From Eqs. (25) and (26), the optimal price point p^* and package size b^* can be found (More details will be covered in Section 4).

The bulk-selling pricing model is just one of the retail pricing strategies. Is it possible to introduce an upfront fee for further VM price reduction? The question leads to "two-part tariff" pricing model, which is also known as the reserved pricing model.

3.7. Reserved pricing model

Reserved (or two-part tariff) pricing model can be considered a price mixing strategy. It consists of two parts of pricing. This model is widely adopted by many service industries, such as retail, entertainment, airline, and telco. The purpose of this model is to give CSPs more flexibility to target various market segments. We can define the model in equations from 27 to 32.

$$CS_i\left[q, p, F\right] = \left(\sum_{j=1}^{q} U_i\left[j\right]\right) - qp - F, \quad 0 < F \leq F_m \tag{27}$$

$$q_i\left(p, F\right) = arg\ \max_{q}\left(CS_i\left[q, p, F\right] \geq 0\right) \tag{28}$$

$$\begin{aligned} &IF\ CS_i\left[q, p, F\right] > 0, \qquad cq_i\left[p, F\right] = 1, \\ &ELSE\ cq_i\left[p, F\right] = 0 \end{aligned} \tag{29}$$

$$Q\left(p, F\right) = \sum_{i=1}^{S} q_i\left(p, F\right) D_i\left[p\right], \tag{30}$$

$$\begin{aligned} &C\left[Q\left(p\right)\right] = c_u\left[Q\left(p, F\right)\right] Q\left(p, F\right) \\ &\pi\left[p, F\right] = pQ\left(p, F\right) - C\left[Q\left(p\right)\right] \\ &\qquad\qquad + F\sum_{i=1}^{S} cq_i\left[p, F\right] D_i\left[p\right] \end{aligned} \tag{31}$$

$$\begin{aligned} &c_u\left[Q\left(p, b\right)\right] \leq p \leq M, \\ &\left[p^*, F^*\right] = \underset{p,\ F}{\operatorname{argmax}}\ \pi\left[p, F\right] \end{aligned} \tag{32}$$

where p^*, F^* are the optimal price for usage charge and optimal reserved fee respectively and F_m is the maximum fee can be estimated. In this case, we set to \$100. $cq_i\left[p, F\right]$, it represents the reserved account for a particular customer in the market segment "i". If the customers' surplus value is less than and equal to zero, it means customers will not pay upfront fee F ($cq_i\left[p, F\right] = 0$).

In comparison with bulk-selling, reserved pricing also has two variables. It means that the cloud customers have to pay the upfront reserved fee, and then they can provision VMs. In return, CSPs offer a significant discount of the usage charge to encourage cloud customers to consume more. The benefit of this model can boost sales and increase profit. If a CSP would like to increase the profit further, the next logical step is to combine both bulk-selling and reserved together.

3.8. Reserved plus bulk-selling

This model is to leverage both bulk-selling and two-part tariff models advantages. However, the benefits of the two models do not have an additive effect. The net profit increase is not bulk-selling plus reserved. Very often, the profit margin increment is small or declining because the cloud customer surplus value may approach its upbound limit when we model them separately. We can use the following equations from 33 to 39 to represent this model.

$$\begin{aligned} &IF\ B = q - b\left\lfloor \tfrac{q}{b} \right\rfloor > 0 \rightarrow CS_i = -200, \\ &\forall b \in q = \{1, 2, \cdots, q_m\}, \qquad q\left(b\right) = nb, \qquad n = 1, 2, \cdots \end{aligned} \tag{33}$$

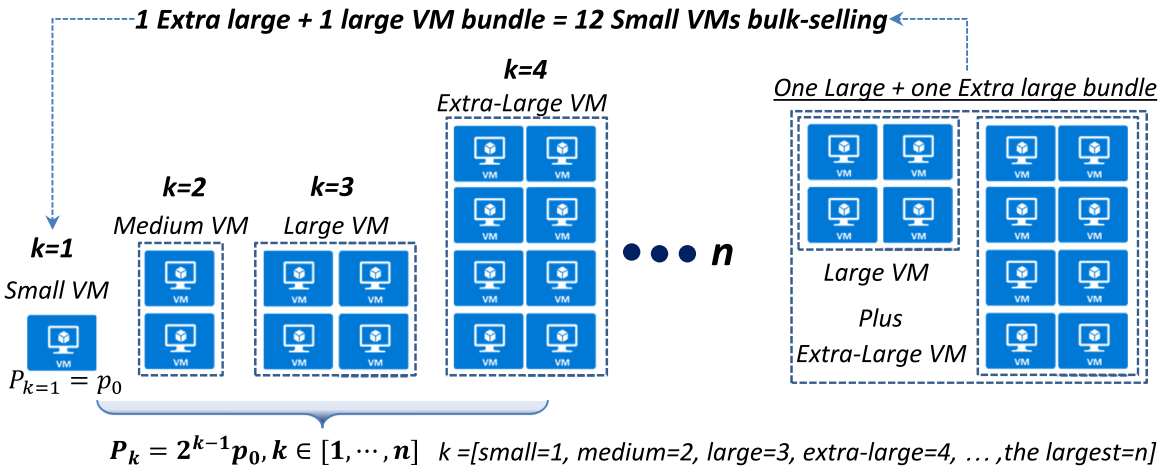

Fig. 5. Cloud service bundle Vs. bulk-selling pricing model.

Otherwise,

$$CS_i[p, q(b), F] = \left(\left(\sum_{j=1}^{q} U_i[j]\right) - pq(b) - F\right), \tag{34}$$

$$0 < F \leq F_m$$

$$q_i[p, b, F] = arg\max_{q}(CS_i[p, q(b), F] \geq 0) \tag{35}$$

$$Q(\cdot) = Q(p, b, F) = \sum_{i=1}^{S} q_i(p, b, F) D_i[p] \tag{36}$$

$$IF\ CS_i[p, q(b), F] > 0$$
$$cq_i[p, b, F] = 1, \quad ELSE\ cq_i[p, b, F] = 0 \tag{37}$$

$$\pi[p, b, F] = pQ(p) - C[Q(p)] + F\sum_{i=1}^{S} cq_i[p, b, F] D_i[p] \tag{38}$$

$$c_u(Q(p)) \leq p \leq M, \qquad C[Q(p)] = c_u(Q(p))\, Q(p)$$
$$[p^*, b^*, F^*] = \underset{p,b,F}{\text{argmax}}\ \pi[p, b, F] \tag{39}$$

The goal of this model is to maximize the profit with bulk-selling to encourage the customers to buy more VMs and with a reserved fee to motivate the cloud customers to consume more for less unit cost per VM. In comparison with other models, this model has three variables to be optimized. Now, the question is how to optimize pricing variables for profit maximization, which the question has been left unanswered in Section 3.3.

4. Genetic algorithm and experiment implementation

4.1. Proposed methods

There are many possible optimization methodologies or techniques that we could apply for the optimizing problem, such as gradient descent, Genetic Algorithm (GA), and simulated annealing. Gradient descent cannot be applied because the profit equation is noncontiguous. Simulated Annealing could be one of the possible methods for our problem because it usually is better than greedy algorithms, but the technique can be slow, especially if the cost function is expensive to compute. Subsequently, we can adopt GA to solve our problem, and we can solve the problem quickly (30 second/per each iteration if an objective function has no further improvement).

4.2. Genetic Algorithm (GA)

The useful properties of GA are (1) It does not require specifying sub-functions explicitly, (2) The objective function can be either differentiable or non-differentiable, (3) It takes less computational memory, (4) It can optimize multiple variables in parallel, and 5.) Some local optimal solution could bring some insights as for potential price options to form an alternative pricing strategy. The basic idea of evolution computation strategy is "trial and error" is shown in Figs. 6 and 7. The principle of this method is based on the underlying microevolution of both mutation and natural selection [3], which is to mimic the biological process that is searching for an optimal solution in a problem domain.

Based on Eqs. (3) and (4), our goal is to find the maximum value of the profit "π" by searching for the optimal price point "p". We know that price, cost, and sales quantity are interdependent. It is challenging to define a precise sub-function for the solution. However, we can set up price p as "genes" and let a set of price, quantity, and unit cost to be a chromosome (a set of parameters for the solution). A string of chromosomes is known as the genome. The entire combination of prices (genes) is known genotype, and the corresponding profits are referred to phenotype, as shown in Fig. 6. Note that the optimal variables can be extended to bulk-selling size "b" or upfront fee "F". In Figs. 6 and 7, we only show both optimal price and bulk size.

In the following example of the on-demand pricing model, we set the price value in the range [0, \$1.5] because no customers will expect to provision the VM more than the maximum amount of their utility value. If we first trial the initial price value or a gene as p =\$0.265, we should have the profit π [p]= \$86, unit cost $c_u[Q[0.265]] = 0.255$, and the total sales quantity Q =9,062. Clearly, it is not an optimal price. So, we let GA compute Eq. (4).

For each of 100 population size (A "standard GA" parameter of the population size can be set up between 100 and 200 [69]), we will keep the best 7.5% of prices p or genes and discard 92.5% in term of better profit values because we set up the mutation rate is 7.5% in our experiment process. After "y" times of this iteration, we find the maximum value of profit based on the performance of the convergence resolution or stopping condition for GA is either $r_{con} = 0.01\%$ Eq. (40) or time out = 30 s (roughly between 280–350 GA iterations) (Fig. 7).

$$r_{con} = \left|\frac{\pi[p_{m+1}] - \pi[p_m]}{\pi[p_m]}\right| < 0.01\%, \qquad m = 1, \cdots, N \tag{40}$$

where $\pi[p_m]$ is profit estimated at iteration m with price p_m.

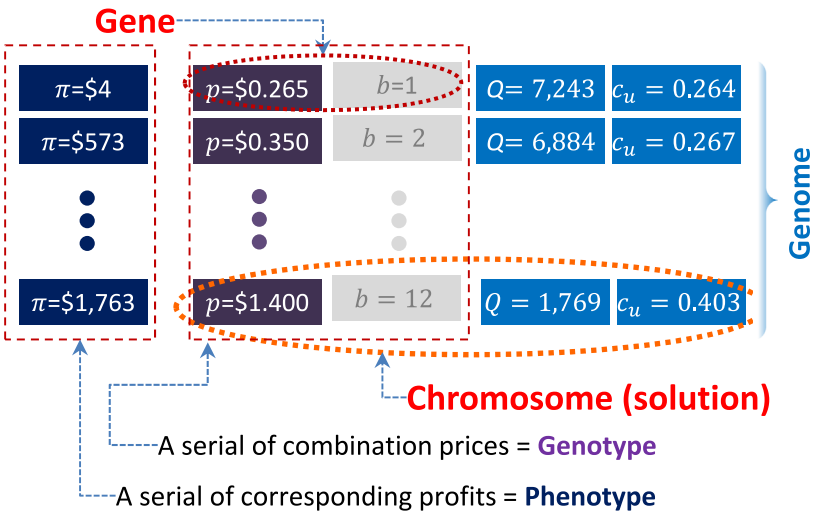

Fig. 6. Details of GA calculation for maximum profit π for on-demand price mode.

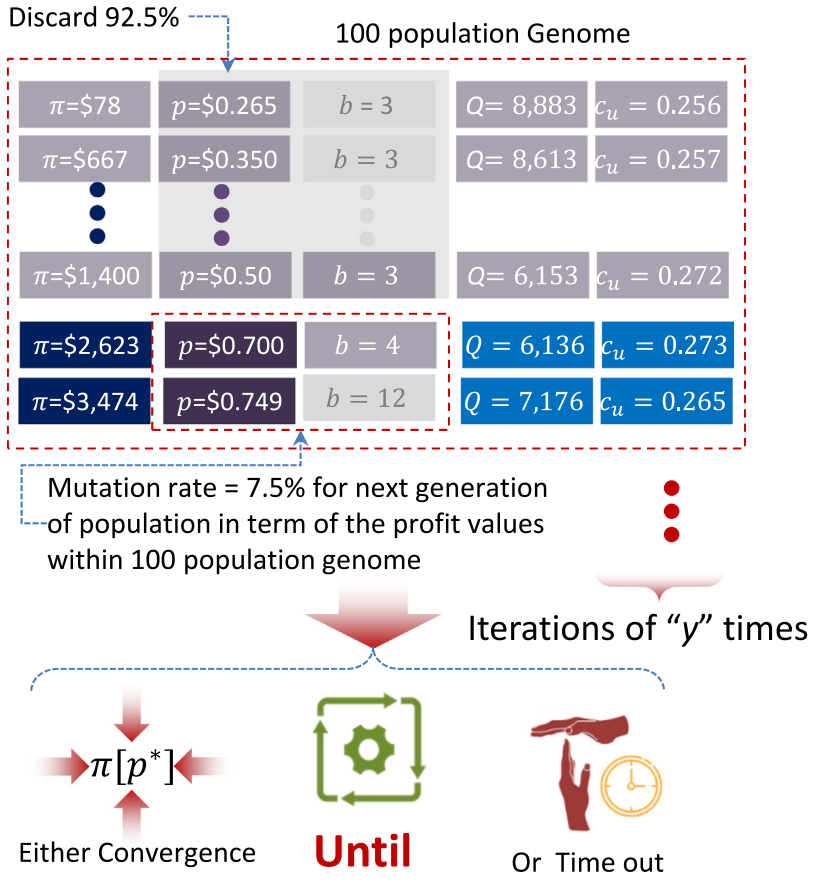

Fig. 7. Performance function and criteria of the GA solution.

4.3. Experiment implementation and pseudocode

A Pseudocode is presented to articulate this genetic algorithm process as Algorithm 1. To carry out this iterative process, we can adopt different software applications to implement our experiments, such as Matlab, R and even Microsoft Excel Solver. R has two convenient packages: GA and Genalg, which can quickly run our experiments. The input data of our experiments are sourced from Table 6 as initialized parameters. The outputs are the optimal values of four pricing models for on-demand, bulk-selling, reserved, and bulk + reserved.

5. Experiments results

5.1. On-demand pricing model results

Table 7 shows the final results for all pricing models that are including on-demand, which CSPs should charge $0.749 per VM instance/hour for the maximum profit of $2,463. The average unit cost is about $0.281. In comparison with cost-based pricing, the on-demand pricing can boost a 66% profit margin if we take account of the external rationality. Although the profit margin (100%) of the cost-based pricing looks very attractive, it is not

Algorithm 1: Pseudocode of Cloud Pricing Models

PROGRAM: Genetic Algorithm for CloudPricingModel

Input: **PopulationSize** $N(m \leftarrow 1 \cdots 100)$,

PriceRange $R \in [0, K_i]$,
CrossoverRate $C_r \leftarrow 0.6$,
MutationRate $m_r \leftarrow 0.075$ // Initalize Parameters;
Output: $p^* \leftarrow arg\max_p \pi[p]$
// Find the Optimal Price p for Cloud Business Profit Maximization;
1. $P_0\{p_m\} \leftarrow \{p_1, p_2 \cdots p_m\} \in [0,\ K_i]$ **InitalizePopulation**
// Randomly Select p_m : Population size, Problem Size;
2. $\pi[p] \leftarrow Q[p] * (p - c_u)$ **Objective Function**
// Calculate Objective Function;
3. $\frac{\pi[p_{m+1}] - \pi[p_m]}{\pi[p_m]}$ **EvaluationPopulation**
// Use Fitness Function for Evolution Initial Population $P_0\{p_m\}$;
4. $\pi[p] \leftarrow p$ **GetBestSolution**
// Assign the Best Price to the Object Function from Initial Population $P_0\{p_m\}$;
5. **While** != **StopCondition** (Time ≤ 30 sec without improvment) OR ($r_{con} <$ 0.01%) **DO**
// either Time Less Than 30 secs or $\pi[p]$ Convergence
6. $P_g\{p_m\} \leftarrow P_0\{p_m\}$ **SelectParents**
// $P_g\{p_m\}$ // Select Parents Population;
7. $cg \leftarrow 0$ **SetToZero** // Sign Children Generation to Zero;
8. **FOREACH** $P_g1, P_g2 \in P_g$ **DO** // Iteration Process
9. $P_{cg}1, P_{cg}2 \leftarrow$ **Crossover**$(P_g1, P_g2,\ C_r)$
//Perform Crossover and Sign to Children Population;
10. $P_{cg} \leftarrow$ **Mutation** $(P_{cg}1, m_r)$ //Perform Mutation;
11. $P_{cg} \leftarrow$ **Mutation** $(P_{cg}2, m_r)$ //Perform Mutation;
12. **ENDFOR**
13. **EvaluatePopulation** (P_{cg})
// Use Fitness Function to Evaluate Children Population P_{cg};
14. $\pi[p] \leftarrow p$ **GetBestSolution** (P_{cg})
// Assign the Best Price to the Object Function from Children Population P_{cg};
15. $P_{cg} \leftarrow P_{cg}$ **Replace** (Population, P_{cg}) //Insert Offspring;
16. $cg \leftarrow cg + 1$ // Create New Generation;
17. **ENDWHILE**
18. **Return** $p^* \leftarrow arg\max_p \pi[p]$;

optimal. The result of this comparison means the cost-based pricing is significantly underestimated the unit price of cloud customers' willingness to pay in this case.

In Fig. 8, we show how the profit is evolving in term of offering prices. There are a few local-optimal prices, such as $0.40, $1.225 $1.350. To overcome these local-optimal values, we can try different initial values of prices or change the parameters of GA. As we indicated in Section 3.5, the on-demand pricing model is just one of the price models for various market segments. Other models, such as reserved or bulk-selling, are possible for CSPs to gain a higher profit.

5.2. Bulk-selling model results

The bulk-selling price model needs to optimize two variables. One is a VM price, and the other is a bulk-selling size. Based on the equations from 21 to 26, the cloud customers will only make a purchase decision when their surplus values are higher than their cost (CSP's offering price). Our experiment results show that the package size is 12 and the optimal price is just slightly below the on-demand price or $0.745 for CSP to achieve the maximum profit margin of 175%. This price will not attract customers to buy in bulk. Subsequently, CSP can give about 7% discount off the on-demand price, which is to set the selling price at $0.70 and reduce the package size from 12 to 4. Even so, the CSP can still achieve a 157% profit margin. If we keep the package size is 4 and give a 7% discount off the on-demand pricing, we can plot out the profit evolution along with the price changing, as shown in Fig. 9.

If we keep the 7% discount price unchanged and make the variation of the bulk-size from 1 to 12, we can find bulk-size-4 is the local optimal and bulk-size -12 is the global-optimal value. These optimal price points provide more price options (as shown in Fig. 10) for CSP to form a better pricing strategy based on its own business environment.

Intuitively, the downside of the bulk-selling is that some cloud customers do not want to scarify their flexibility of Pay as You Go (PAYG) and look for a competitive price because their business might not require a bulk-size of VMs. As a result, customers might switch to other cloud competitors. Adopting one price model could cause a CSP to lose some market share when the CSP insists on some pricing models, such as bulk-selling model. If a CSP would like to keep both higher profit margin and market share, what is an alternative?

5.3. Reserved price model results

The possible solution is a reserved pricing model. Our experiment result shows that the reserved price model can achieve a profit margin of 203%. The main profit contribution is due to the reserved fee, which is $5.701 per account or $3,410. The VM price is $0.273, which is very close to the unit cost, which $0.272. If we keep the reserve fees the same ($5.701) and changing the VM price, we can see the cloud price evolution (See Fig. 11).

Again, if the VM price is kept the same (at $0.273 per VM) and the reserved fee is changed, there will be two local-optimal prices at $1.50 and $7.35 shown in Fig. 12.

5.4. Bulk plus reserved results

If CSPs would like to increase profit further, they can combine both bulk-selling and reserved models. In comparison with a pure reserved model, "bulk + reserved" can grow only about 2% profit. This model offers different alternatives for CSPs to form a pricing strategy to meet various requirements in different market segments, which a CSP can increase the usage charge and decrease the reserved fee or vise verse. The plot of profit, sales' volume and unit cost along with VM price change can be considered as a combined effect of bulk-selling plus reserved as observed in Fig. 13.

Following a similar principle, we can also plot the fee change while the unit price ($0.6597) and bulk size (12) are kept the same. The result is shown in Fig. 14. As we should see, the shapes of the two plots are very similar except the sales volume.

Overall, our experimental results show that the on-demand pricing model can significantly increase CSPs profit margin in comparison with the cost-based pricing. The bulk-selling price model is aiming to encourage customers to buy more for less usage charge. The reserved pricing model is to decrease more usage charge with the upfront reserved fee. This flexible option can help CSP to maintain a healthy profit margin while the usage price is kept very competitive. The "bulk + reserved" model is to provide different options of cloud pricing strategies to maximize the CSP's profit while they can target various cloud market segments.

6. Analysis and discussion

This study demonstrates a comprehensive framework of how to formulate four value-based cloud pricing models from a customer's value co-creation perspective. In contrast to previous works that assumed a uniform market with only one utility function, this solution of cloud pricing is much realistic and practical because market segmentation practice has been widely applied to many service industries. Cloud industry is not exceptional. AWS has adopted up to seven different types of pricing models (spot, on-demand, reserve, bare-metal, dedicated host, and Code on Demand) for different market segments. Based on multiple market segments, we can leverage the GA to find the optimal pricing solution for each model.

Table 7
The result of on-demand pricing model.

Pricing models	Type of pricing models	Optimal price: p	Optimal Bulk Size	Reserved Fee *F*	Unit Cost: c_u	Total Cost C	Total Sales Quantity: q	Total Revenue: R	Maximum Profit: π	Profit Margin
Markup	Cost-based	$0.557	NA	NA	$0.278	$1,538	5,525	$3,076	$1,538	100%
On-demand	Value-based	$0.749	NA	NA	$0.281	$1,479	5,256	$3,942	$2,463	166%
Bulk-Selling	Value-Based	$0.700	4	NA	$0.273	$1,672	6,136	$4,295	$2,623	157%
Reserved Fee	Value-based	$0.273	0	$5.701	$0.272	$1,676	6,152	$5,085	$3,410	203%
Bulk+Reserved	Value-based	$0.660	12	$1.084	$0.265	$1,901	7,166	$5,382	$3,481	183%

Note: Bulk-selling price is based on 4 VMs per package of bulk.

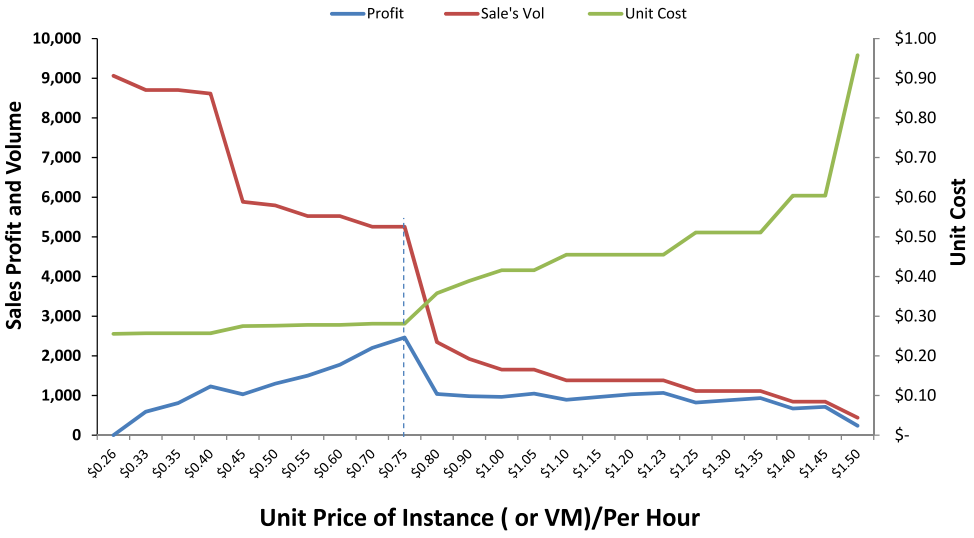

Fig. 8. On-demand price model of price change for optimizing profit, sales volume. and unit cost.

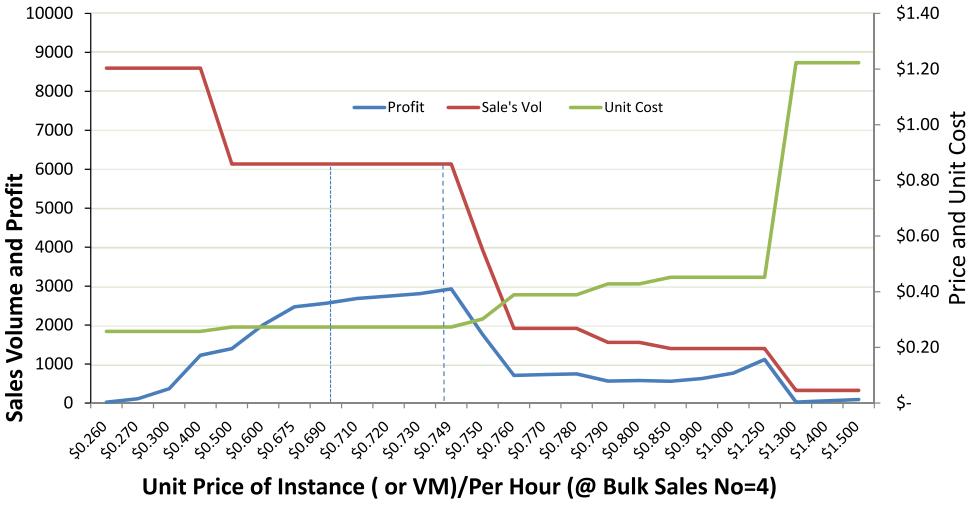

Fig. 9. Bulk-selling price models of price change for optimized profit, revenue, sales vol., and unit cost (BulkSize@4).

6.1. GA performance evaluation

In comparison with other optimal solutions, the GA process requires less computing memory and power and does not need to specify sub-functions. The object function does not have to be differentiable. It can be either continuous or discrete. Many software packages can implement the GA process. Even MS Excel Solver can implement it, which is very handy for many practitioners to generate pricing options and form a better and competitive pricing strategy. The GA process can also be updated quickly if the cloud market environment has been changed.

To evaluate the performance of the GA process for the optimal pricing value, we tune one of the GA's parameters: mutation rate into different values and to see which we can achieve a better performance result quickly. According to [70,71], we applied the mutation rates between 0.001 and 0.5. Our result shows when the mutation rate is equal to 0.075, the profit margin of reserved pricing models can be quickly converged to the maximum value (as shown in Fig. 15) within the specified timeframe of 30 s with 100 population size and converged rate of 0.01%.

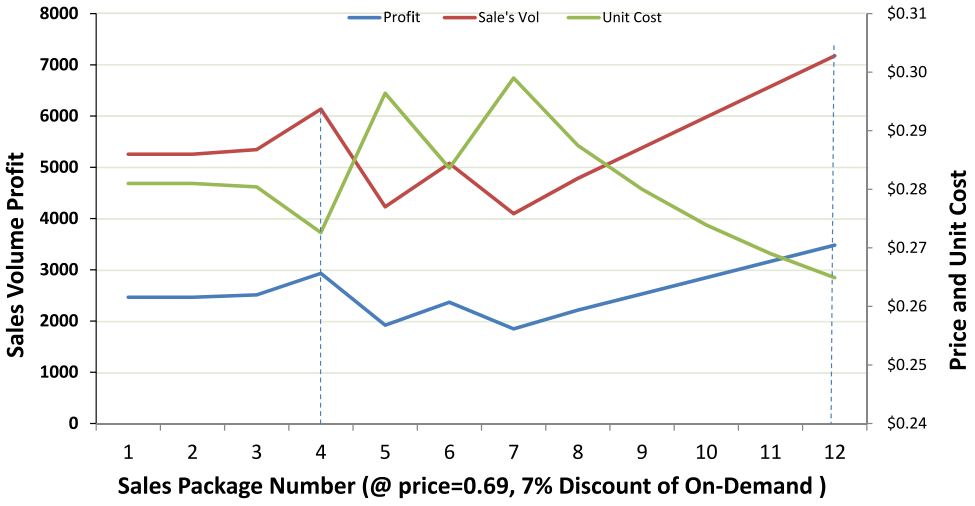

Fig. 10. Bulk-selling package size evolution.

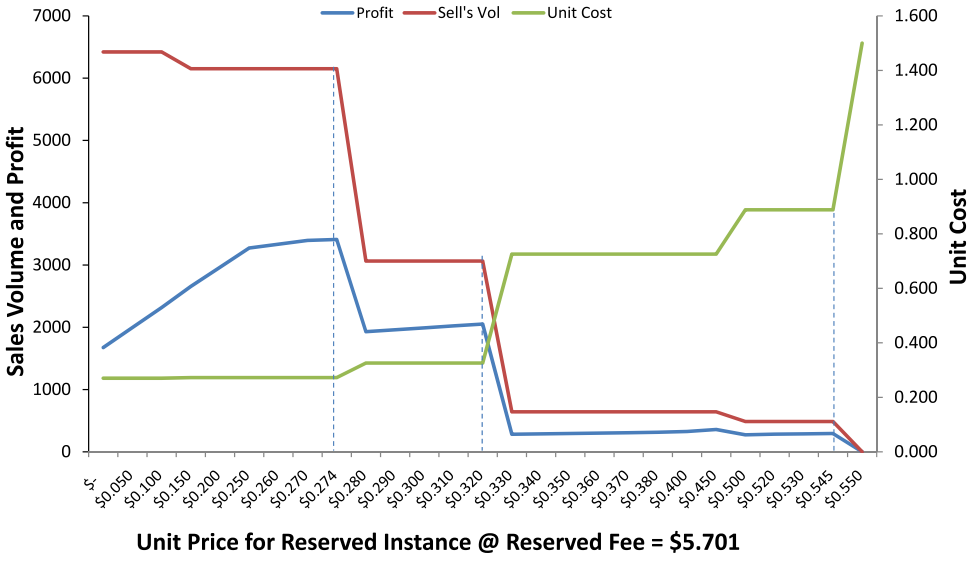

Fig. 11. Reserved pricing model of price change for optimized profit, sales vol., and unit cost @ F = $5.701.

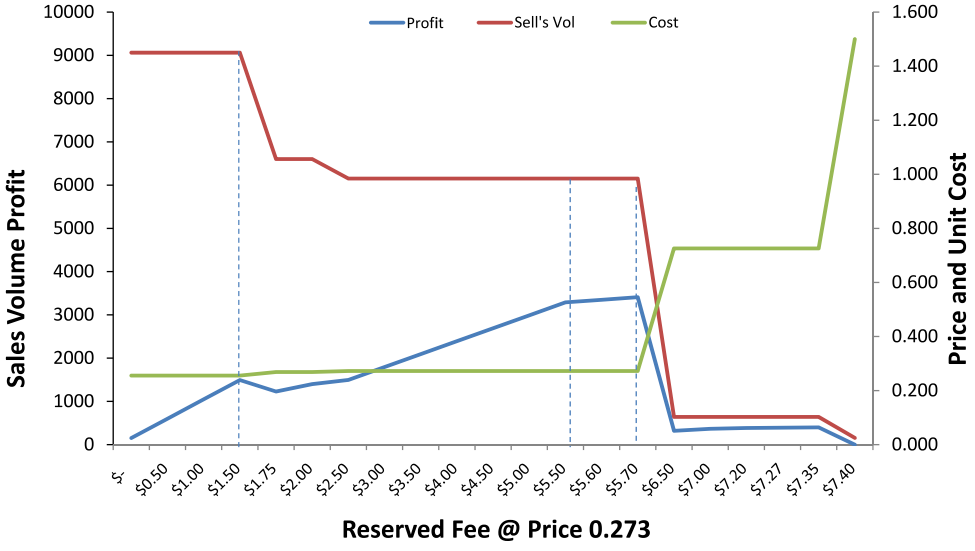

Fig. 12. Optimal reserved fee change for optimized profit, revenue, sales volume, and unit cost (Price@$0.273).

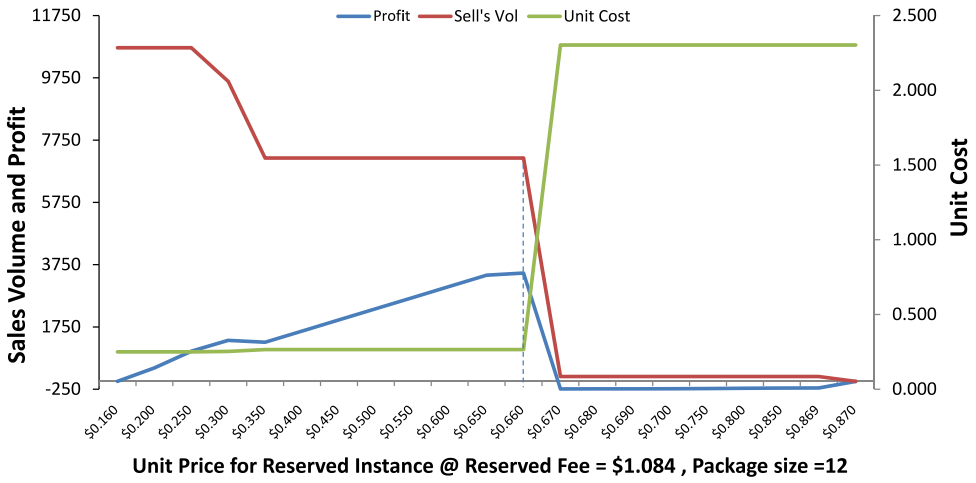

Fig. 13. VM price change of reserved + bulk for optimized profit, sales volume, unit cost, (F @$1.084 bulk size@12)

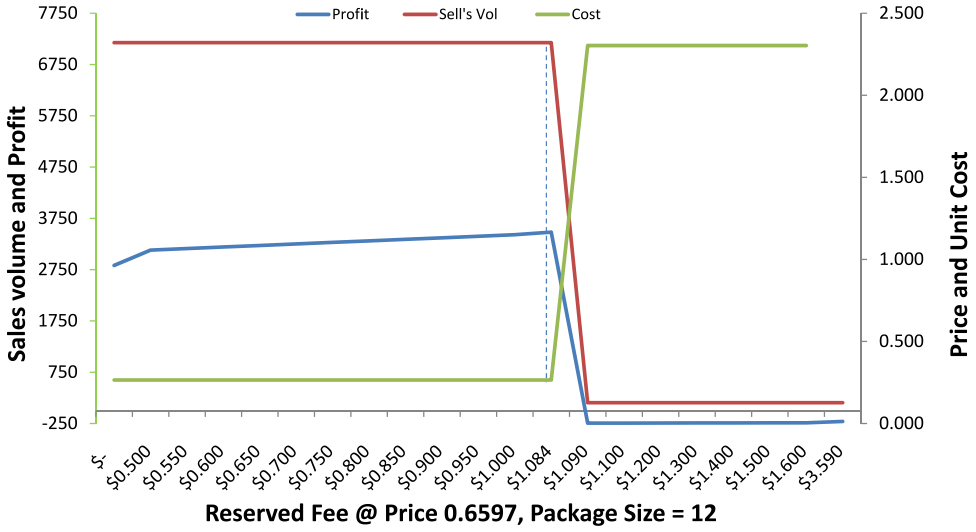

Fig. 14. Fee change o bulk + reserved for optimized profit, sales volume, and unit cost (Price @$0.6597, bulk size @12)

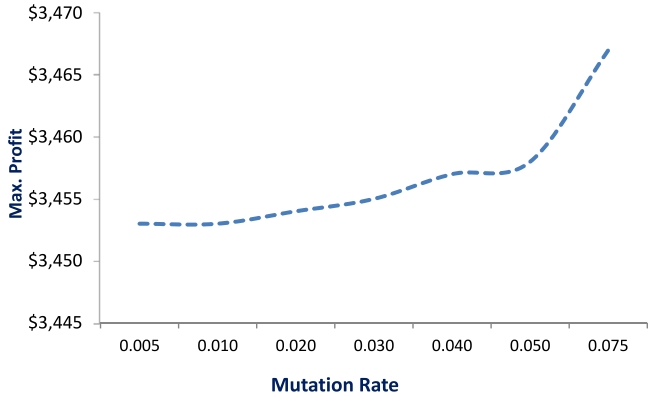

Fig. 15. GA performance evaluation for different mutation rate.

6.2. Comparison with created pricing models

From our experiment results, we illustrated that cost-based pricing has the lowest profit gain ($1625). As Nagle et al. [2] indicated, although the model carries a financial legitimacy, it only provides "mediocre financial performance". Although a 100% profit margin seems to be very attractive, it is "mediocre". The critical issue of cost-based pricing is that it excludes external rationality

On the other hand, on-demand can achieve a higher profit margin and higher sales volume in comparing with cost-based pricing. However, the on-demand model might work well with one business application (or one market segment), but not fit with others. To solve this issue, we have created both bulk-selling and reserved pricing models for different cloud applications. One of the advantages of bulk-selling and reserved models is that they can provide business certainty for cloud resource capacity. The downside is that cloud customers could lose some flexibility. If we compare all four cloud pricing models, the reserved + bulk selling pricing model can achieve the highest profit gain ($3481) for CSP, as shown in Fig. 16.

In order to gain a higher profit margin, bulk-selling price model is one of the good pricing strategies, which we have observed in the cloud pricing practice. In fact, bulk-selling is equivalent to AWS, Azure, IBM Cloud, and Google Cloud Platform's reserved instance (without an upfront fee). The only difference is time. With bulk-selling, cloud customers have to provision all resources at the same time. In contrast, AWS, Azure, IBM cloud, or GCP's reserved instances can be consumed from one or three years. The longer time of cloud resource reservation, the cheaper

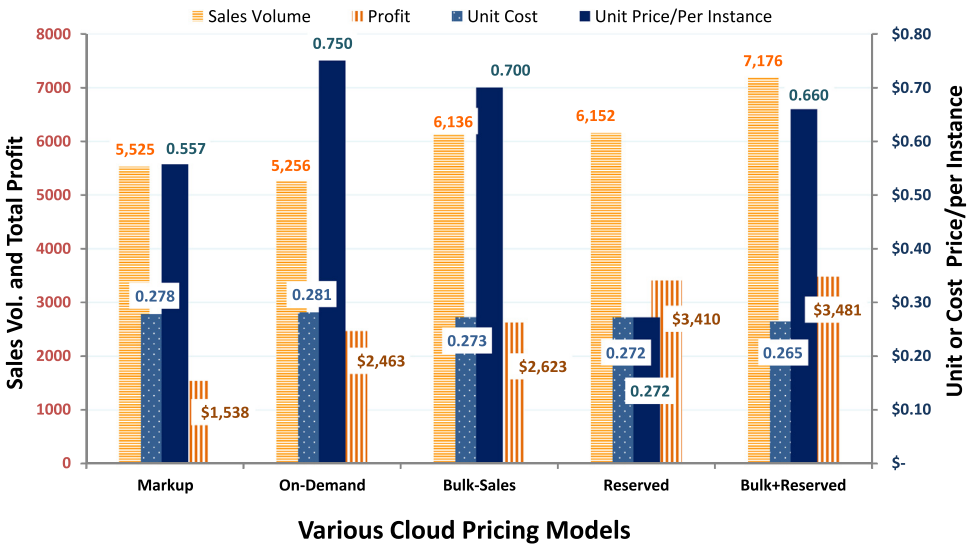

Fig. 16. Comparison of different pricing models.

unit price is. The reservation time is equivalent to a bulk-size. If we put time and cloud assets depreciation factors aside, the currently reserved instances offered by major CSPs are similar to the bulk-selling or bundle pricing model. As we showed in our experiments, the bulk-selling price model can improve CSP's profit by 6% even with even a 7% price discount in comparison with the on-demand price. (See Table 7 and Fig. 16.) That is why many CSPs encourage cloud customers to adopt bulk-selling with a discount price.

If CSPs would like to make further improvement of their profit margin, they can introduce the upfront reserved fee (two-part tariff) and reduce the usage charge of VM unit price in return. With the upfront reserved fee, the CSPs can reduce VM usage charge as low as the production cost and still maintain a healthy profit margin, which is around 203%. Comparing with the "on-demand" model, the usage charge (or unit price of VM) drops nearly 64%. Now, the upfront reserved fee becomes the major profit contributor to CSP's profit. If the cloud customers are not willing to pay a higher upfront reserved fee, CSPs can adopt the mixing model of bulk + reserved fee. The above simulation result shows that by a combination of the bulk and upfront reserved fee, CSPs can lift profit growth and reduce the upfront fee by 81% (in comparison with pure reserved model) and decrease VM price by 12% (in comparison with on-demand).

On the sales volume of VM across all segments, we can find customers of segment 2 would not purchase any number of VM for proposed price models, as shown in Table 8. This is because their utility function is risk-taking. The shape of the utility function is concave. None of the above-proposed pricing models would capture customers' surplus values in segment 2 unless a CSP can offer a substantial discount, such as 50%–55% price reduction of the on-demand, which is similar as spot, preemptible, and low priority pricing model.

Having a considerable price discount for one market segment alone and scarifying other markets' values is not a good pricing strategy because the cloud business profit will decline significantly. For example, if the price is dropped by 40% across all market segments, the profit margin will be reduced by about 58%. Selling cloud service with a substantial discount price is not a sustainable business practice for CSPs.

However, what we have observed is that many leading CSPs do offer a discount price, such as spot instance, preemptible and low-priority, for risk-taking customers. The reason that CSPs can offer a massive discount without cannibalizing the profits from other pricing models is that the cloud service with a discount price has many restricted conditions, such as preemptible, time limit, limited availability zone, legacy infrastructure, etc.

From a marketing perspective, the spot or preemptible instance is more like "Razor-and-Blades" pricing strategy [72], which is to use a lower price to simulate customer's demand. Practically, it is not a good idea for business customers to rely on spot or preemptible instances (VMs) alone for their mission-critical application, although the price of spot instance is very competitive.

Table 8 shows that the customers in segment 5 will purchase more than what they need if a CSP offers bulk-selling price models only. However, if all prices models are offered to various customers spontaneously in a cloud market, the customers of segment 1 will provision VMs according to bulk-selling price model because it has the highest surplus value and the lowest cost, which we assume the lowest unit price as the purchasing decision criteria. Customers of segment 2 will purchase nothing. Customers of segment 3 will select the on-demand model. Customers of segment 4 will also choose bulk-selling. Customers of segment 5 will be the same as segment 3. Customers of segment 6 will prefer bulk-selling pricing model. The sum of six market segments for all value-based price models is also 40. The average profit margin is over 161%, and the total cost for CSP is just slightly increasing by 0.72% in comparison with cost-based pricing. (Note: if a discount rate of bulk-selling is changed, the customers of the 6th market segment will favor of reserved pricing model)

These value-based price models provide a wide range of pricing options for CSP to achieve the maximum profit by capturing more customers' surplus values from various market segments. Based on the market segmentation theory [1], the ideal strategy for CSPs is to have personalized pricing because of the better the information about the customers, the fine partition of the customers into a group and the larger the possibilities for CSP to extract customer surplus". Theoretically speaking, the ideal solution is that one price is dedicated to one customer, which is also known as the 1st order price discrimination. However, it would be impossible for CSP to implement personalized pricing strategy because it requires a lot of managerial and sales' resources. The alternative solution is "market segmentation". Naturally, different market segments will lead to different utility values. It results in various price models with multiple optimal price points to meet different preferences. Table 9 provides summary information of

Table 8
Sales volume of VM for each model.

Pricing Models	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	Total
Cost-based	7	0	2	12	6	12	39
On-demand	6	0	**2**	12	**6**	12	38
Bulk-Selling	**8**	0	4	**12**	8	12	44
Reserved Fee	10	0	0	12	6	**12**	40
Bulk + Reserved	12	0	0	12	12	12	48
All Models	**8**	**0**	**2**	**12**	**6**	**12**	**40**
Quantity	1,614	0	360	2,820	648	156	5,598
Max. Revenue	$1,210	$0	$270	$1,974	$486	$109	$4,050
Cost	$454	$0	$101	$769	$182	$43	$1,549
Max. Profit	$756	$0	$169	$1,205	$304	$67	$2,501
Preferred price Model	Bulk	none	On-Demand	Bulk	On-Demand	Reserved	

Table 9
Summary of all pricing models.

Pricing strategy	Pricing models	Application scenarios	Advantages	Disadvantages
Cost-Based	Cost-Based Pricing	Enterprise internal cost modeling	Recover the cost bottom line	Arbitrary
Value-Based	On-Demand	Application develop, solution architecture	Flexible	High Cost
	Bulk-Selling	Long-term Web hosting required large server cluster, Deliver SLA App.	Having a certain % price discount	Have to buy in a bulk size
	Reserved	Having cloud resource certainty	Relative lower cost	Lack of flexibility
	Bulk + Reserved	Large server clustering & resource certainty	Min. cost	Lowest flexibility

Table 10
Profit, Revenue and Optimal Price Comparison with Other Works.

Sources	Six segments	Uniform market [43] [44] [31]	Uniform market [28]	Uniform market [46]	Uniform market [45]	Uniform market [30]
Utility	Six utility functions	Iso-Elastic utility, α<1	Iso-Elastic utility, $\alpha = 1$	Iso-Elastic utility, α>1	Linear by diminish return	Exponential utility
Equivalent utility function	$U_i(q), i = 1 \cdots 6$	$U(q) = K\frac{q^{1-\alpha}}{1-\alpha}$	$U(q) = K$	$U(q) = K\frac{q^{1-\alpha}}{1-\alpha}$	$U(q) = U_0 - \alpha p$	$U(q) = K(1-e^{-\alpha q})$
Optimal cost	$0.28	$0.399	$0.336	$0.604	$0.376	$0.259
Optimal price	$0.749	$0.957	$0.750	$1.499	$0.954	$0.415
Max. Profit	$2,463	$1,008	$1,044	$757	$1,202	$1303
Max. revenue	$3,942	$1730	$1,936	$1,267	$1,983	$3,460
Profit loss	0%	59%	58%	**69%**	51%	**47%**
Revenue loss	0%	56%	51%	**68%**	50%	**12%**
Sales vol.	5,256	1,808	2,581	845	2,077	5,884

all models that we have proposed in this work in term of different application scenarios, advantages, and disadvantages.

6.3. Comparison with other works

To the best of knowledge, there has been no research work to propose multiple market segments for Cloudonomics. Although some previous works [31] claimed that the uniform price would not suffer any revenue loss in comparison with the 1st order price discrimination, we illustrated this claim was contradicting to the theory of market segmentation [1,15]. Based on our experiment result, we have demonstrated that if we assume there is uniform market defined by either iso-elasticity or linear or exponential utility function, the profit loss will be from 47% up to 69% and the revenue loss will be from 12% up to 68% shown in Table 10.

In summary, our value-based price modeling, together with the comprehensive pricing framework, is better than the current state of the art of cloud price modeling, which has been highlighted in Table 4.

7. Conclusions and future work

This study has developed an overall framework of the pricing process that is how to generate various price models and how to find these optimal price points of each model for CSP to maximize the profit. These are two elements of pricing strategy (Shown in Fig. 1) that have been demystified in our research work. The significance of this study is that it presents a comprehensive and practical process for value-based pricing.

We demonstrate how to establish four types of practices price models, which are known as on-demand, bulk-selling, reserved, and bulk-selling + reserved pricing models. While the modeling process appears to maximize CSP's profit, it is actually a value co-creation because the modeling process is to generate a partnership with cloud business customers. This modeling process becomes a practical tool for any CSP to construct their cloud price models based on the defined business strategy, cloud market environment, and their expertise.

We show how to use the GA to find the optimal price points by maximizing CSP's profit. Our experiment results demonstrate that the reserved pricing model can achieve the best profit margin,

which is about 203% while the bulk-selling is the most favorite model if a 7% discount rate of the on-demand is applied. It implies that the single pricing model with an assumption of a uniform market does not necessarily mean it can achieve the maximum profit for CSPs. Our simulation results reiterate the importance of cloud market segmentation.

The results also illustrate that our proposed models could not capture the customers who have risk-taking utility, which often belongs to a niche market segment. The only discount price model can satisfy the customers who are willing to take a high risk for cloud resource uncertainty. If a CSP wants to capture the surplus value of this niche market, the CSP should carefully design a particular price model not only to target the niche market segment but also isolate it and avoid the discount pricing model to cannibalize the higher profit margin from other cloud market segments. On the other hand, CSP should not always select the price model that can generate the highest profit margin only because there are many competitors in the cloud market. Consequently, our future work will extend from a monopoly market assumption to oligopolies or competitive cloud market environment.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Claycamp, W. Massy, A theory of market segmentation, J. Mark. Res. (JMR) 5 (4) (1968) p. 388–394.
[2] T.T. Nagle, G. Müller, The Strategy and Tactics of Pricing: A Guide To Growing more Profitably, Pearson, Boston, Mass, 2011, pp. p. 1–13.
[3] R. Leardi, Nature Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier, 2003.
[4] G.N. Mankiw, Principles of economics, Cengage Learn. (2014) p. 425 & p. 447.
[5] V. Ramaswamy, K. Ozcan, What is co-creation? An interactional creation framework and its implications for value creation, J. Bus. Res. 84 (2017) p. 196–205.
[6] M. Kohtamaki, R. Rajala, Theory, and practice of value co-creation in B2B systems, Ind. Mark. Manage. 56 (2016) p. 4–13.
[7] A. Hinterhuber, Customer value-based pricing strategies: Why companies resist, J. Bus. Strategy 29 (4) (2008) p. 41–50.
[8] R. Grewal, G.L. Lilien, Handbook of Business-To-Business Marketing, Edward Elgar Pub., Northampton, Mass, 2012, p. 3–12.
[9] M. Ikefuji, R.J. Laeven, J.R. Magnus, C. Muris, Pareto utility, Theory Decis. 75 (2013) p. 43–57.
[10] C. Wu, R. Buyya, K. Ramamohanarao, Cloud computing market segmentation, in: Proceedings of the 13th International Conference on Software Technologies (ICSOFT 2018), 2018, p. 888–897.
[11] C. Wu, R. Buyya, K. Ramamohanarao, Modeling cloud business customers' utility functions, 2019, http://www.buyya.com/papers/Cloud-Utility-Functions.pdf (Accessed 10 June 2019).
[12] D. Yankelovich, D. Meer, Rediscovering market segmentation, Harv. Bus. Rev. 84 (2) (2006) p. 122–131.
[13] https://github.com/google/cluster-data/blob/master/TraceVersion1.md (Accessed 20 August 2018).
[14] https://www.amd.com/Documents/AMD_WP_Virtualizing_Server_Workloads-PID.pdf (Accessed 20 August 2018).
[15] M. McDonald, I. Dunbar, Market Segmentation How to Do It and How to Profit from It, John Wiley & Sons, 2012, p. 16.
[16] O. Michalski, S. Demiliani, Implementing Azure Cloud Design Patterns, Packt Publishing, Birmingham, 2018, p. 109–119.
[17] C. Wu, R. Buyya, Cloud Data Center Cost Modeling, a Complete Guide To Planning, Designing and Building a Cloud Data Center, Morgan Kaufmann, San Francisco, CA, USA, 2015.
[18] V. Tarasov, D. Hildebrand, G. Kuenning, E. Zadok, Virtual machine workloads: the Case for new benchmarks for NAS, Fast (2013) p. 307-320.
[19] A.K. Mishra, J.L. Hellerstein, W. Cirne, C.R. Das, Towards characterizing cloud backend workloads: Insights from google compute clusters', Perform. Eval. Rev. 4 (2010) p. 34.
[20] G.A. Jehle, P.J. Reny, Advanced Microeconomic Theory, Pearson Education Edinburgh Gate, Harlow, 2011, p. 4 & p. 288.
[21] J. De Koning, Service design geographies, in: Proceedings of the ServDes. 2016 Conference, No. 125, Linköping University Electronic Press, 2016.
[22] M.C. Calzarossa, L. Massari, D. Tessera, Workload characterization: A survey revisited, ACM Comput. Surv. 48 (3) (2016) p. 48.
[23] A. Undheim, P. Heegaard, Differentiated availability in cloud computing SLAs, in: 2011 IEEE/ACM 12th International Conference on Grid Computing, on, 2011, p. 129.
[24] M. Zhang, Utility function in autonomic workload management for DBMSs, Int. J. Adv. Intell. Syst. 5 (1 & 2) (2012).
[25] M. Maurer, V.C. Emeakaroha, L. Brandic, J. Altmann, Cost–benefit analysis of an SLA mapping approach for defining standardized cloud computing goods, Future Gener. Comput. Syst. 28 (1) (2012) p. 39–47.
[26] A. Williams, M. Arlitt, C. Williamson, K. Barker, Web Workload Characterization: Ten Years Later, 2005, p. 3–21.
[27] P. Belleflamme, M. Peitz, Industrial organization markets and strategies, 2011, p. 27, p. 139.
[28] A.N. Toosi, K. Vanmechelen, K. Ramamohanarao, R. Buyya, Revenue maximization with optimal Capacity control in infrastructure as a service cloud markets, IEEE Trans. Cloud Comput. 3 (3) (2015) p. 261–274.
[29] H. Xu, B. Li, Dynamic cloud pricing for revenue maximization, IEEE Trans. Cloud Comput. 1 (2) (2013) p. 158–171.
[30] M. Aazam, H. Eui-Nam, M. St-Hilaire, L. Chung-Horn, I. Lambadaris, Cloud customer's historical record based resource pricing, IEEE Trans. Parallel Distrib. Syst. 27 (7) (2016) 1929–1940.
[31] H. Xu, B. Li, A study of pricing for cloud resources, ACM Sigmetrics Perform. Eval. Rev. 40 (4) (2013) p. 3-12.
[32] O. Agmon Ben-Yehuda, M. Ben-Yehuda, A. Schuster, D. Tsafrir, Deconstructing amazon EC2 spot instance pricing, ACM Trans. Econ. Comput. 1 (3) (2013) p. 16.
[33] I. Hirose, J. Olson, The Oxford Handbook of Value Theory, Oxford University Press (OUP), 2015, p. 13.
[34] J. Altmann, M.M. Kashef, Cost model-based service placement in federated hybrid clouds, Future Gener. Comput. Syst. 41 (2014) p. 79–90.
[35] M. Macías, J. Guitart, A genetic model for pricing in cloud computing markets', in: Proceedings of the 2011 ACM Symposium: Applied Computing, 2011, p. 113.
[36] S. Novani, in: Kuntoro Mangkusubroto, et al. (Eds.), Value Co-Creation on Cloud Computing: A Case Study of Bandung City, Indonesia, Systems Science for Complex Policy Making: A Study of Indonesia, in: Translation Systems Sciences, vol. 3, Springer Nature, New York, 2016, p. 43–63.
[37] C. Kilcioglu, J. Rao, Competition on Price and Quality in Cloud Computing, WWW 2016, April 11–15, Montréal, Québec, Canada ACM, 2016.
[38] C.S. Yeo, S. Venugopal, X. Chu, R. Buyya, Autonomic metered pricing for a utility computing service, Future Gener. Comput. Syst. (2010) p. 1368–1380.
[39] J. Sherwani, N. Ali, N. Lotia, Z. Hayat, R. Buyya, Libra: a computational economy-based job scheduling system for clusters, Softw. Pract. Exper. (6) (2004) p. 573.
[40] C.S. Yeo, R. Buyya, Pricing for utility-driven resource management and allocation in clusters, Int. J. High-Performance Comput. Appl. (4) (2007) p. 405.
[41] J.N. Franklin, Methods of Mathematical Economics: Linear and Nonlinear Programming, Fixed-Point Theorems, SIAM, 2002, p. 190.
[42] P. Hande, M. Chiang, R. Calderbank, J. Zhang, Pricing under constraints in access networks: Revenue maximization and congestion management, in: INFOCOM, 2010 Proceedings IEEE, 2010, pp. 1–9.
[43] C. Joe-Wong, S. Soumya, Mathematical frameworks for pricing in the cloud: Revenue, fairness, and resource allocations, 2012, CoRR arXiv:abs/1212.0022.
[44] M. Shahrad, C. Klein, L. Zheng, M. Chiang, E. Elmroth, D. Wentzlaff, Incentivizing self-capping to increase cloud utilization, in: Proceedings of the 2017 Symposium on Cloud Computing, September 24, ACM, 2017, pp. 52–65.
[45] J. Chen, C. Wang, B.B. Zhou, L. Sun, Y.C. Lee, Tradeoffs Between Profit and Customer Satisfaction for Service Provisioning in the Cloud, in: Proceedings of the 20th International Symposium: High-Performance Distributed Computing, 2011, p. 229.
[46] F. Alzhouri, A. Agarwal, Dynamic pricing scheme: Towards cloud revenue maximization, in: IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, 2015, p. 168.
[47] L. Du, Pricing and resource allocation in a cloud computing market, in: Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid 2012), 2012, p. 817.
[48] E. Brynjolfsson, P. Hofmann, J. Jordan, Cloud computing and electricity: Beyond the utility model, Commun. ACM 53 (5) (2010) 32–34.
[49] https://moz.com/blog/crawl-outage (Access 20 August 2018).

[50] I.A. Kash, P.B. Key, Pricing the cloud, IEEE Internet Comput. Internet Comput. (1) (2016) 36–43.
[51] N. Jain, A truthful mechanism for value-based scheduling in cloud computing, Theory Comput. Syst. 54 (3) (2014) 388–406.
[52] C. Grönroos, Quo Vadis, Quo vadis marketing? Toward a relationship marketing paradigm, J. Mark. Manage. 10 (5) (1994) 347–360.
[53] C. Wu, A.N. Toosi, R. Buyya, K. Ramamohanarao, Hedonic pricing of cloud computing services, IEEE Trans. Cloud Comput. Cloud Comput. (2018) 1–15.
[54] B. Rady, Serverless Single Page Apps, Fast, Scalable and Available, Pragmatic Bookshelf, June 14, first ed., 2016, pp. 3–4.
[55] P. Krugman, R. Wells, Economics Fourth Edition, Worth Publishers, New York, 2015, pp. 282–283.
[56] https://www.asbfeo.gov.au/sites/default/files/Small_Business_Statistical_Report-Final.pdf (Accessed 20 August 2018).
[57] P. Wakker, Explaining the characteristics of the power (CRRA) utility family, Health Econ. Chichester 12 (2008) 1329.
[58] O.C. Ibe, Markov Processes for Stochastic Modeling, Elsevier, Amsterdam, Netherlands, 2013, pp. 55–82.
[59] M. Fahrioglu, F.L. Alvarado, Using utility information to calibrate customer demand management behavior models, IEEE Trans. Power Syst. (2) (2001) 317.
[60] U. Bhat, An Introduction To Queueing Theory: Modeling and Analysis in Applications, Birkhäuser, Boston, 2010, pp. 34–40.
[61] A. Homer, J. Sharp, L. Brader, M. Narumoto, T. Swanson, Cloud design patterns: Prescriptive architecture guidance for cloud applications, Microsoft Patterns Pract. (2014) 150.
[62] X. Tang, J. Xu, S.T. Chanson (Eds.), Web Content Delivery Web Information Systems Engineering and Internet Technologies Book Series, Vol. 2, Springer, Boston, MA., 2005.
[63] P. Bats, Scalability, and economics of citrix XenApp and citrix XenDesktop 7.6 on Amazon web services, 2014, https://www.citrix.com/content/dam/citrix/en_us/documents/partner-documents/scalability-and-economics-of-citrix-xenapp-and-citrix-xendesktop-76-on-amazon-web-services.pdf (Accessed 28 August 2018).
[64] F. Schimscheimer, Workload considerations for virtual desktop reference architecture, 2018, https://www.vmware.com/techpapers/2009/workload-considerations-for-virtual-desktop-refer-10081html (Accessed 28 August 2018).
[65] P.H. Nakhai, N.B. Anuar, Performance Evaluation of Virtual Desktop Operating Systems in Virtual Desktop Infrastructure, in: 2017 IEEE Conference on Application, Information and Network Security (AINS), Application, Information and Network Security (AINS), 2017 IEEE Conference on, 2017, p. 105.
[66] P.J. Davis, E. Garcés, Quantitative Techniques for Competition and Antitrust Analysis, Princeton University Press, Princeton, 2010.
[67] S. Landsburg, Price Theory and Applications Minneapolis/St, Paul: West Pub. Co., 1995.
[68] https://www.rightscale.com/blog/cloud-cost-analysis/aws-vs-azure-vs-google-cloud-pricing-compute-instances, (Accessed 30 August 2018).
[69] S. Rylander, B. Gotshall, Optimal population size and the genetic algorithm, Population 100 (400) (2002) p. 900.
[70] K. Matthias, S. Thomas, S. Horst, Variable mutation rate at genetic algorithms: introduction of chromosome fitness in connection with a multi-chromosome representation, Int. J. Comput. Appl. 72 (17) (2013).
[71] J. He, L. Guangming, Average convergence rate of evolutionary algorithms, IEEE Trans. Evol. Comput. 20 (2) (2016) p. 316–321.
[72] R.C. Picker, The razors-and-blades myth(s), Univ. Chicago Law Rev. (1) (2011) p. 225.

Caesar Wu is a senior IEEE and ACM member and has just completed his Ph.D. journey at the University of Melbourne recently. He is the first author of Cloud Data Center and Cost Modeling. He was a senior IT and network design engineer in Telstra for nearly 20 years. He designed, built managed, and operated many of Telstra's enterprises and IT data centers. He has over 30 years working, researching and academic experiences across various industries.

Rajkumar Buyya is a Redmond Barry Distinguished Professor and Director of the CLOUDS Laboratory at the University of Melbourne. He has authored more than 600 publications, and four textbooks are recognized as a "Web of Science Highly Cited Researcher" both in 2016 and 2017 by Thomson Reuters, a Fellow of IEEE, and Scopus Researcher of the Year 2017 with Excellence in Innovative Research Award by Elsevier for his outstanding contributions to Cloud computing.

Ramamohanarao(Rao) Kotagiri received Ph.D. from Monash University in 1980. He has been at Melbourne University since 1980 and was appointed as a professor in 1989. Rao was Head of CS and SE and Head of the School of EE and CS. He received Distinguished Contribution/Service Awards from ICDM, PAKDD, DASFAA, etc. Rao is a Fellow of the Institute of Engineers Australia, Fellow of Australian Academy Technological Sciences and Engineering and Fellow of Australian Academy of Science.