Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges

CAESAR WU, RAJKUMAR BUYYA, and KOTAGIRI RAMAMOHANARAO, Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Victoria 3010, Australia

This article provides a systematic review of cloud pricing in an interdisciplinary approach. It examines many historical cases of pricing in practice and tracks down multiple roots of pricing in research. The aim is to help both cloud service provider (CSP) and cloud customers to capture the essence of cloud pricing when they need to make a critical decision either to achieve competitive advantages or to manage cloud resource effectively. Currently, the number of available pricing schemes in the cloud market is overwhelming. It is an intricate issue to understand these schemes and associated pricing models clearly due to involving several domains of knowledge, such as cloud technologies, microeconomics, operations research, and value theory. Some earlier studies have introduced this topic unsystematically. Their approaches inevitably lead to much confusion for many cloud decision-makers. To address their weaknesses, we present a comprehensive taxonomy of cloud pricing, which is driven by a framework of three fundamental pricing strategies that are built on nine cloud pricing categories. These categories can be further mapped onto a total of 60 pricing models. Many of the pricing models have been already adopted by CSPs. Others have been widespread across in other industries. We give descriptions of these model categories and highlight both advantages and disadvantages. Moreover, this article offers an extensive survey of many cloud pricing models that were proposed by many researchers during the past decade. Based on the survey, we identify four trends of cloud pricing and the general direction, which is moving from intrinsic value per physical box to extrinsic value per serverless sandbox. We conclude that hyper-converged cloud resources pool supported by cloud orchestration, virtual machine, Open Application Programming Interface, and serverless sandbox will drive the future of cloud pricing.

CCS Concepts: • **General and reference** → *Surveys and overviews*;

Additional Key Words and Phrases: Cloud services provider (CSP), cloud price model, value-based pricing, market-based pricing, cost-based pricing

ACM Reference format:
Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao. 2019. Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges. *ACM Comput. Surv.* 52, 6, Article 108 (October 2019), 36 pages.
https://doi.org/10.1145/3342103

Authors' addresses: C. Wu, R. Buyya, and K. Ramamohanarao, Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Victoria 3010, Australia; emails: {caesar.wu, rbuyya, rkotagiri}@unimelb.edu.au.

0360-0300/2019/10-ART108 $15.00
https://doi.org/10.1145/3342103

1 INTRODUCTION

The cloud computing transformation is now gaining momentum [1, 2]. It has entered the "early majority" of the cloud technology adoption lifecycle, where cloud computing has become a mainstream market of the IT infrastructure [4]. According to Wikibon [3], the Compound Annual Growth Rate (CAGR) of a true private cloud (a hyper-converged cloud solution) alone will grow 29.2% from 2017 to 2027, while IaaS will grow 15.2% during the same period. However, one critical issue has still been ambiguous, which is how to comprehend a variety of cloud pricing models that are offered by different Cloud Service Providers (CSPs) systematically. Yet, the number of pricing schemes in the current cloud market is overwhelming. The aim of this article is to provide a systematic view of many pricing models for both CSPs and cloud customers[1] so that CSPs can be competitive and achieve sustainability, while cloud customers can make the best decisions during the cloud transformation.

Recently, many CSPs or cloud computing advocators claim that cloud computing is cheaper computing due to its Total Cost of Ownership (TCO) [5, 6]. However, Weinman [8] argued that "Cloud Computing is not cheap computing." Martens et al. [9] echoed this view, and he noticed that many cloud cost (price) conclusions lack a systematic approach in the real costs estimating behind various cloud pricing models. Many favored claims are often dependent on ad-hoc processing of price modeling without the consideration of many indirect and hidden factors.

As a result, Buyya et al. [10, 11] suggested that the topic of cloud computing pricing should be considered in an interdisciplinary way, which should be studied under the scope of multiple disciplines, including cloud technologies, price theory, microeconomics, operations research, and value theory. Similarly, Kash and Key [102] also indicated that "current cloud pricing schemes are fairly simple." "Multidimensional scheduling and pricing offer greater potential for increasing both customer satisfaction and (CSP)'s revenue" with a growing number of new cloud service features. According to References [12, 13], no single discipline can provide a satisfying solution for cloud pricing. An isolation approach of cloud pricing could increase the difficulty for decision-makers to comprehend the benefits and risks of cloud services as well as a price to be paid. One of the examples is how to understand Amazon Web Services (AWS) spot instance or spot block (up to 6-hour service duration time) pricing. It can be considered as dynamic-based pricing[2] because of the nature of fluctuation influenced by supply and demand [37, 38]. However, it can also be regarded as auction-based, cost-based, or time-based pricing due to its multiple characteristics [97, 98]. Therefore, we argue that the cloud pricing issue must be examined by its value propositions and an interdisciplinary approach.

Although we draw multiple disciplines for cloud pricing, we mainly focus on four knowledge domains: The cloud pricing model is the focal point. Microeconomics is our theoretical tool to understand the cloud price that is influenced by supply and demand in the cloud market place. Value theory is the measurement for a customer's value proposition, because "we do not know the meaning of a (value) concept unless we have a (theoretical) method of measurement for it" [135]. Operation research is a method to help cloud decision-makers make better decisions for any given cloud price during the cloud transformation. Cloud technologies allow CSP to create innovative cloud service features along with new pricing models to capture maximum customer surplus values from

[1] Cloud business customers have their own business, such search engine optimization (SEO), storage backup, virus scanning and and so on., run on the cloud infrastructure to serve other customers. They are not end users. From a cloud customer's perspective, CSP's cloud price is equivalent to its cost.

[2] The dynamic pricing model means the price is a function of many variables, such as time, season, customer demand, and so on. Many firms adopt this price to manage their yield for their limited capacity or resources. It has been widely applied in many service and utility industries such as airline, hotel, electric and gas utilities.

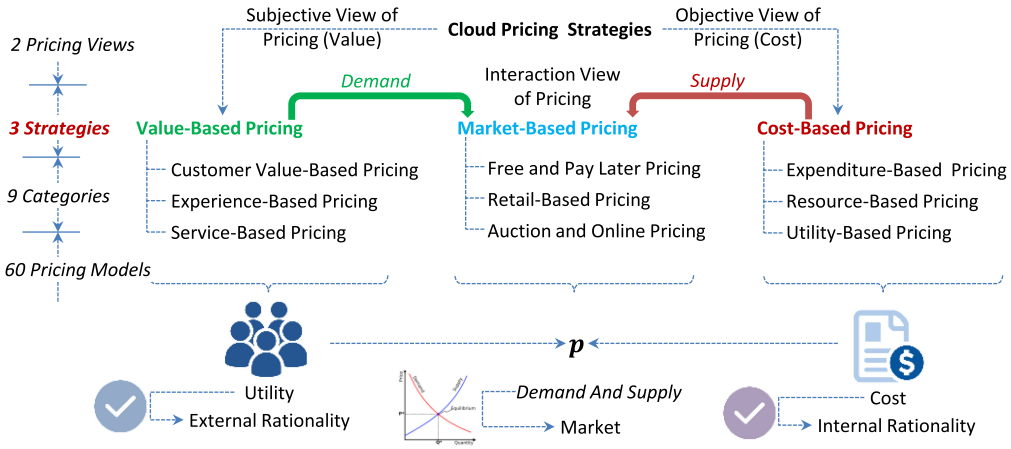

Fig. 1. A hierarchical framework of Cloud Pricing Strategies.

multiple cloud market segments. Throughout this taxonomy and survey, we will examine both the pros and cons of different cloud pricing strategies[3] and model categories[4] regarding a fundamental question of value [99], which is represented by various subjective experiences of many cloud business customers. These subjective experiences are often measured by Cloud Service Metrics (CSM) [103, 104], such as acquisition, retention, and efficiency from a business customer's perspective.

Overall, we derive three strategies of cloud pricing through both subjective (values) and objective (fact) views. Value-based pricing is demand driven, and cost-based pricing is supply driven. Moreover, market-based pricing can be seen as the result of an equilibrium of both supply and demand in a cloud market. Based on these basic strategies, we can define a hierarchical pricing framework that is illustrated in Figure 1. Each layer of the framework is driven by its goal. At the top, the pricing is driven by the principle of value [99]. The next layer down is derived from three pricing strategies, which are to pursue a long-term goal of the business. The layer further down is drawn from pricing tactical,[5] which is oriented by short-term objects. The aim of tactical pricing is how to translate a pricing strategy into tactical objects. Finally, the bottom layer of cloud pricing consists of 60 individual models, which is detail oriented. It explains the details of implementing a pricing strategy. This framework implies that if a strategy is cost based, then the final price "p" is determined by a cost that is driven by internal rationality. In contrast, if a strategy is value based, "p" is dependent on cloud customers' utility value, which is determined by external rationality. If a strategy is market-based pricing, then "p" is a result of the market equilibrium of supply and demand. The essence of this hierarchical framework is to reflect the microeconomics [16] in term of price theory.

Based on this framework of categories, we can find that many earlier works mainly focused on either cost-based or market-based pricing and paid less attention to value-based pricing. Therefore, this study will include all three pricing strategies and pay special attention to value-based pricing. We argue that even if a CSP knows all the pricing components (facts) of a cloud (such as cloud

[3]Strategy is how does a decision maker deal with or solve the given business problem for a long term or overall goal.
[4]Model is a representation of strategy. It can help us to visualize and access the relationship of the various objects. It is a simplified or abstracted description of reality, especially a mathematical one, for us to predict the future.
[5]Tactic is similar as a strategy, which is a plan to achieve a specified aim. However, the aim of a tactic is to gain immediate or short-term benefits rather than long term one. It is possible to win a game tactically but lose it strategically. Many tactics can support an overall strategy.

service cost, markup ratio, market share and target rate of return, etc.) [14, 15] objectively, then a cloud price still cannot be determined, because a decision-maker does not know how to handle these facts, such as which item (fact) is more important than the others and why and when it is much more important than others. These questions are a question of value [19]. If we would insist to derivate from value alone, then it becomes a naturalistic fallacy [20], which, as Nagel et al. [17] demonstrated, this kind of pricing strategy would become absurd. To avoid this logic fallacy, this work will provide a comprehensive framework by considering both CSP's cost and customers' value proposition for cloud pricing. As a result, we made the following contributions:

- We categorize 60 pricing models into three pricing strategies and nine pricing categories. Many models have not been considered by CSPs yet, but they have been widely adopted by other service industries, such as airline, travel, hotel, recreation, healthcare, telecom, and retail sectors. The purpose of revealing these potential models is to help many CSPs to compete on pricing, not on a price.
- We reveal most of the recently proposed models in considerable depth regarding their contributions and gaps plus their business application. Moreover, our work also highlights characteristics of pricing models offered by leading CSPs, and they often leverage their business strength to build their models.
- We identify four research challenges of cloud pricing: (1) how to move from pure cost-based to both value-based and cost-based pricing, (2) how to move from statefulness to stateless[6] resource pricing, (3) how to transfer from mutable to immutable[7] pricing, and (4) how to develop the cloud pricing models to capture more cloud customers surplus values[8] along with the cloud infrastructure lifecycle and new technology eruption.
- We also provide some preliminary ideas on how to approach these challenges in principle.

The rest of the article is organized as follows: Section 2 reviews the history of cloud pricing from a practical perspective. It includes cloud service launch times and virtualization technologies that underpin different cloud prices and cloud business. The aim of having this historical overview is to understand the multiple roots of cloud pricing models proposed by many researchers during the past decade. We then outline three pricing strategies based on value theory. Section 3 establishes the taxonomy of cloud pricing models. Section 4 provides a detailed survey of selected papers that were published from 2008 to the present. Finally, we compare each pricing model with other models for its methodology and theoretical roots. Section 5 provides our conclusions and four possible development trends in cloud pricing. Based on these trends, we highlight four challenges and possible solutions. All acronyms in this article are listed in Table 1 (Online Appendix A).

2 HISTORY OF CLOUD PRICING MODELS

2.1 Cloud Pricing Models in Practice

The first cloud pricing model can approximately be traced back to Salesforce.com's Russian doll model[9] [100], which are similar to optimal feature pricing (one of the retail-based pricing models,

[6]Statefulness means a backend hosting server or VM maintains user's state information in the sessions form. In contrast, Stateless does not keep any state information for the end-user. Anything is stored on the end-user or client's side in the form of a cache.

[7]It is a programming term, which means the value of some objects (e.g. variable, data structure, a function, or a method) can be altered or updated while the term of immutable means the value of the object cannot be changed.

[8]Surplus Value is also called as consumer surplus. It means a value difference between a price of willingness to pay and the actual price has been paid.

[9]Russian Doll or Matryoshka Doll pricing model is a type of marketing strategy to bundle different product features into one nested deal, like a Russian Doll.

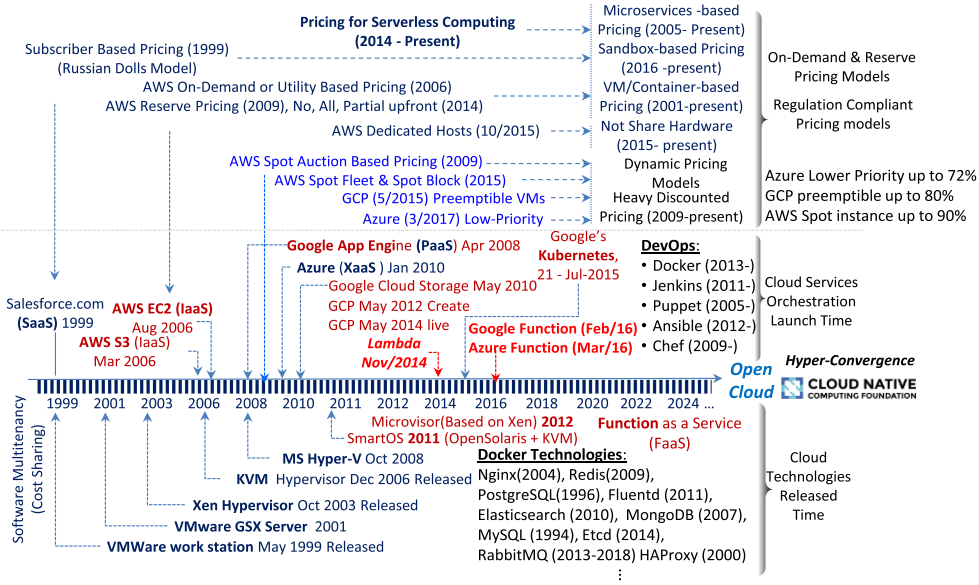

Fig. 2. A short history of cloud service pricing models and enabling cloud technologies.

more details in Section 3.6). Salesforce.com's pricing model is a contrast to Siebel's distributed or perpetual licensing model. In 2000, the average price of Siebel's Customer Relationship Management (CRM) software would be around $10,000 per license plus additional $5,000 ongoing costs for a patch, regular upgrades, bugs fixes, maintenance, backup, and help desk support. Consequently, it is beyond the reach of many small and medium enterprises (SME), because they could not afford to allocate a significant amount of IT budget or Capital expenditure (Capex) upfront. This issue led to an opportunity for Marc Benioff (one of the founders of Saleforce.com) to offer a subscription-based pricing model for SaaS [21].

The cloud technology that underpins subscription-based pricing is software multi-tenancy. The idea of multitenancy is an analog to drawing from an apartment building where the tenants can share costs, such as a public facility, security, and so on, but still have their private space. By the same principle, Microsoft Hotmail or Google's Gmail also offers the email service, which every user (or tenant) can enjoy the email service via a web browser without any stress of installation and configuration of the mail software by themselves. Figure 2 summarizes a timeline of different pricing models that were adopted by some leading CSPs along with cloud technologies development.

Following a similar idea of sharing, AWS adopted the "on-demand" pricing model for its Simple Storage Services (S3), launched in March 2006 and released Elastic Compute Cloud (EC2) in August 2006 for its public cloud. The enabling technology for AWS is Xen hypervisor, and Citrix Systems released the initial version in October 2003. Later in 2009, AWS launched spot instance (Auction or Dynamic-based pricing) with a substantial discount (up to 90%) in comparison with an on-demand price, but it has some restricted conditions for the services. In 2015, AWS started to offer two modified pricing models for spot instance: Spot Fleet and Spot Block. Following AWS's lead, Google App Engine began to offer a cloud service platform (Platform as a Service or PaaS) for its customers to host their web applications within the current Google data centers in 2008. Its price model is very similar to AWS, but GCP's price is charged in per minute base for Pay as you Go (PAYG). The underlying hypervisor of GCP to support its PaaS is Kernel-based Virtual Machine

(KVM) that was initially released by Qumranet[10] in 2006. In 2015, GCP also offered a discount (up to 80%) price or preemptible model for its cloud service to match AWS's spot model. In comparison with both AWS and GCP, Microsoft Azure started its cloud business in Jan 2010. Its price models are very similar to both AWS and GCP. Azure has quickly captured a significant market share, according to Gartner's Magic [7]. Azure's virtualizing technology is built upon its own Hyper–V, which launched in 2008. In 2017, Azure also followed the footstep of both AWS and GCP to offer a "low priority" price model for up to 72% discount rate in comparison with "on-demand."

Although the top three leading CSPs use three different hypervisors, many public CSPs adopt Citrix Xen, such as IBM Softlayer, Rackspace, GoGrid, Oracle VM for x86, Aliyun (Both Xen and KVM), and Virtustream's μVM (or Microvisor). Linode moved its VMs from Xen to KVM in June 2015, because it believes that KVM is 28% faster than Xen. However, the most popular hypervisor for many private clouds is still dominated by VMware, which is the first commercial hypervisor that was launched in 1999.

Some public CSPs, such as CenturyLink and Interoute, also adopt VMware to support their cloud business, because VMware provides a comprehensive toolset that allows customers to manage their private cloud service efficiently. However, some analysts [105, 106] suggest that if host applications are migrated to a public cloud, then it will become too heavy and cumbersome. One of the interesting observations is that most public CSPs adopt Xen hypervisors and the minimum billing unit of on-demand is per hour base. However, if CSPs adopt the KVM hypervisor, then the billing unit is reduced to a per-minute base. Some CSPs that adopt VMware often require customers to have a long-term commitment for their cloud service contract. In general, virtualization technologies allow CSP to cut out the idle time of cloud data centers and improve cloud resource efficiency by 4 to 5 times. It enables CSP to reduce a significant amount of cloud infrastructure footprints. As a result, CSP can offer various competitive cloud prices to its customers. Table 2 (Online Appendix C) highlights the various price models and underlying hypervisors.

From a CSP perspective, we argue that discount pricing models alone would not be possible to support cloud business profitability and sustainability. Instead, on-demand and reserve models are the profit-driving forces for CSPs. The reasons to offer a discount price are as follows: (1) CSP can fully utilize its spare cloud capacity, (2) CSP can manage its cloud resources effectively for its cloud infrastructure lifecycle, (3) it can capture more customers' surplus values at a lower end of the pricing spectrum, (4) it can become one of the marketing campaign tools for CSP to prompt other cloud services, and (5) it can reduce customer churning by combining discount pricing with on-demand. Recently, AWS offered a modified version of spot instance: spot block and spot fleet, which combines on-demand and spot pricing. In comparison with pure on-demand, both models can save typically 30–45% cost plus a further 5% off for a non-peak time in a region. This is an excellent example to illustrate the AWS pricing strategy to reduce customer churning.

From a cloud customer's perspective, the reserved pricing model is to assure cloud resource certainty, and the on-demand pricing model is to accommodate customer's workload fluctuation with advantages of minimum provisioning time and speed to market. Currently, there are at least seven types of mainstream pricing models in the cloud market, namely On-demand, Reserve, Subscription, Discount (including auction), Code on Demand (CoD), bare metal, and Dedicated Host illustrated in Figure 3. These pricing models are mainly driven by cloud customers' utility values and market segments [107]. These models only show the practical aspect of cloud pricing in history. What is the theoretical aspect of cloud pricing in research?

[10]Qumranet was acquired by Red Hat in 2008, but Red Hat was taken over by IBM in later 2018.

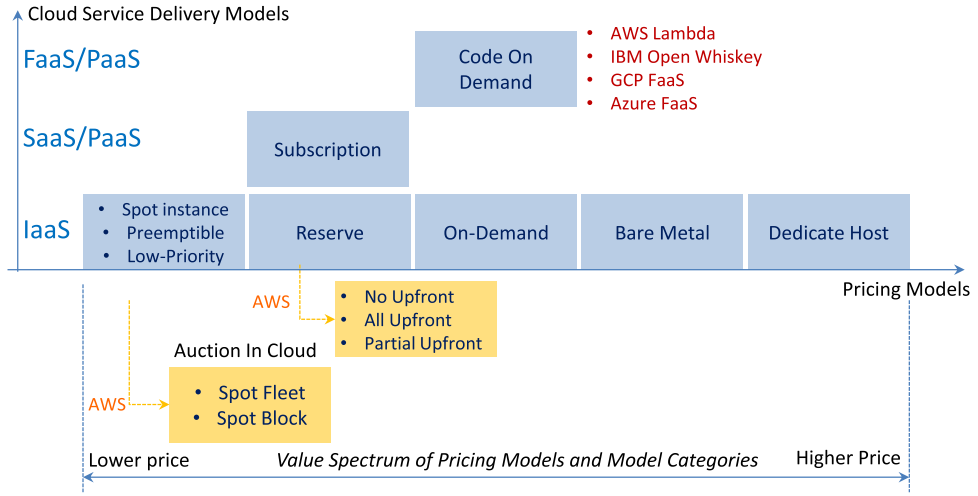

Fig. 3. Summary of cloud pricing spectrum in the current cloud industry.

2.2 Multiple Roots of Cloud Pricing Models In Research

The reason for examining various pricing theories is to clarify multiple roots of cloud pricing theories so that we can make a cloud pricing sense from its theoretical context for a taxonomy purpose. By tracing down the historical roots of various cloud pricing models proposed by many researchers (more details in the following Section 4), we can find the origin of cloud pricing models does not come from single but multiple threads. The current term of cloud pricing model is an amalgam of different sources. According to more than hundreds of research papers from 2008 to now, we can identify possible four primary roots of cloud pricing as shown in Figure 12 (Online Appendix B), which are Utility-computing, Network computing, CSP's profit-driven, and cloud customer performance orientation. This historical tracking suggests two possible options to classify various cloud pricing models. One is to classify pricing models by its historical roots, and the other is to carve (cloud pricing) nature at its (economic) joint [108]. In this article, we will present the taxonomy of the cloud pricing models based on economics and value theory, because it aligns the cloud pricing taxonomy with economic theory and helps many decision-makers to understand business values of each model in term of profit maximization.

2.3 Key Terms, Strategies, and Relationship of Pricing Models

From practice to theory, we have introduced many terms regarding a cloud pricing model. However, the meanings of some key terms and their relationship are still vague in term of cloud pricing contexts, such as price, pricing, pricing scheme, pricing model, pricing structure, pricing category, pricing strategies, value, and customer benefits. These terms and their relationship are essential for the following taxonomy and survey.

The term price is an estimated value or a value tag of cloud service (e.g., $1.00/per hour). Pricing is to give an estimated value based on a value proposition. The pricing scheme[11] is a price plan or cloud service package linked to a pricing tag. It may also be considered as a price configuration. Some CSPs allow cloud customers to create their own pricing scheme by setting a range of standard

[11] AWS c4.larg instance consists of 3.75GB-RAM, 8-ECU or EC2 Compute Unit, 2-vCPU or virtual Central Processing Unit, Linux-OS, and is marked as $0.10/per hour at US East Ohio data center in April 2019.

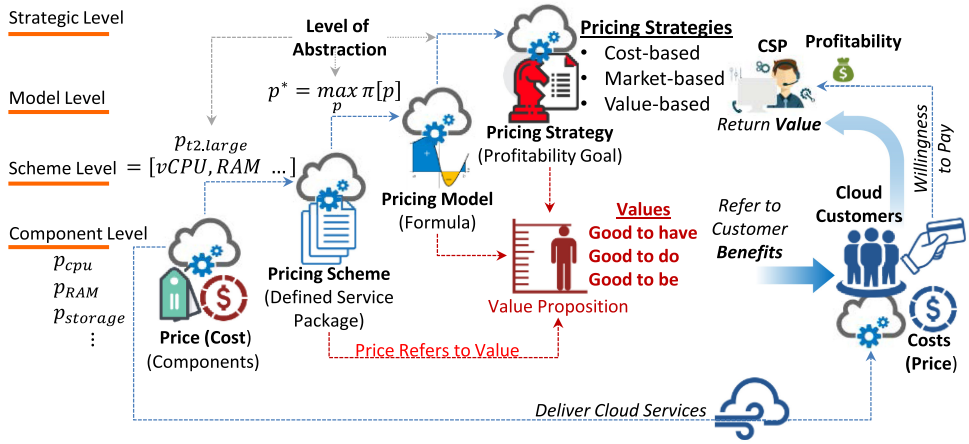

Fig. 4. Relationship of key terms for pricing models, abstraction level, and value proposition.

prices. Pricing model[12] (e.g., on-demand or reserve) is a simplified description that is often defined by a mathematic function for CSP's profit maximization. Pricing category is a group of pricing models has some common characteristics while pricing strategy is an overall plan by coordinating various activities to achieve a long-term business goal.[13] If the pricing scheme is an abstraction of different prices of cloud components, then the pricing model is an abstraction of pricing scheme, and pricing strategy can be considered as an abstraction of pricing model. They are all dependent on a set of value propositions for the purpose of delivering cloud customers' benefits and CSP's profit maximization (see to Figure 4).

The term "value" means how much worth an object has for an agent. It is measured by a unit of utility [122] (worth, satisfaction, and subjective experience), which concerns whetherthings are good or bad in a successful and efficient sense [22, 24]. To this extent, it can be further articulated into three types of good values: (1) "good to have" (e.g., a pricing strategy aims to consolidate good customers' experiences of cloud services), (2) "good to do" (e.g., the strategy drives the customers' value proposition of willingness to pay, which focus on new values), and (3) "good to be" (e.g., a strategy is to simulate customers demand to migrate more workloads to off-premises). The aim of having three types of "good" is to know how to handle all the facts of the pricing model so that a cloud decision-maker can know which fact is more important than others. By delivering "good" values, customers are willing to pay (W2P) for the cloud service benefits, and CSP will get a profit reward from its cloud service delivery. It means "value co-creation" [127]. We can briefly illustrate the relationship all these key terms in Figure 4. Among these terms, pricing strategies are at the top level of abstraction in term of a value proposition. Let us clarify the meaning of cloud pricing strategy.

2.3.1 Value-based Pricing. In comparison with other pricing strategies, value-based pricing is much more subjective. It might not be necessary to reflect on a market price and service costs. A typical example is perception-value, which is based on the customers' perceptions of what is expected compared with what is to be delivered by a CSP. The common term of perceptive value is value for money, that is, the ratio between the worth of a cloud service and a price to be paid [23]. According to Sheth et al. [25], customers perceived values have five dimensions: functional,

[12] $p^* = \max_p \pi[R(p) - C(p)]$, where π is a profit, p is a price, $R(p)$ is total revenue and $C(p)$ is a total cost.
[13] A strategic goal is to achieve a 20% revenue growth in next five years.

conditional, social, emotional, and epistemic values. The final decision of customer choice is a function of multiple perceived values. The main benefit of value-based pricing is that it provides competitive advantages to capture a wide range of cloud services' values [26], such as emotional and epistemic. However, it is quite challenging to be constructed, because "perceived values" are primarily measured by the satisfaction of the individual customer. With the B2B type of service [27], it is even hard to detect end-users' satisfaction directly. Instead, the perceived values could be influenced by an indirect person, such as a manager's or decision maker's perception,

In principle, value-based pricing emphasizes the measurements of customer's experience, satisfaction, and expectation. It includes both intrinsic values,[14] e.g., CPU, RAM, bandwidth, and extrinsic (or instrumental) values,[15] which are determined by the relationship about something, e.g., Pay as You Go (PAYG), 24 × 7 supports, burstable CPU, resource auto-scaling, and so on [113]. Value-based pricing is often applied to innovative cloud service features and some new niche market segments. By a similar line of reasoning, we can extend the value-based criteria to both market-based and cost-based pricing. Consequently, we can construct a 3 × 3 (3 value propositions × 3 pricing strategies) matrix as the classification criteria to differentiate various cloud pricing models listed in Table 3 (Online Appendix D).

2.3.2 Market-based Pricing Strategy. "Market-based pricing" is driven by the equilibrium of all customers and CSPs [28]. "Freemium" is one example of market-based pricing, which it has become popular due to rising FaaS (further details in Sections 3.4 and 4.5). "Freemium" is to give away a product with basic functionality or features for free to gain market share [29].

The primary purpose of Freemium is to convert free customers into premium buyers by giving away just enough value. "Freemium" has been adopted by many CSPs, such as AWS, GoGrid, SoftLayer, Dimension Data, Microsoft Azure, ElasticHosts, and Dropbox. The market-based pricing takes into consideration two kinds of impacts on pricing: price sensitivity and market competitiveness for similar services. Practically, CSP may adopt different pricing models to implement its market-based pricing strategy. Moreover, these models can be measured by various metrics. Marius F. Niculescu et al. [30] highlighted four different measurements: features, quantity, quality, and period. These models can attract many high-end customers and get valuable feedback from a large number of people for a CSP to improve its services.

2.3.3 Cost-based Pricing Strategy. Although market-based pricing is common for many retailer businesses, most of the enterprises and government agents with on-premises cloud infrastructure often adopt cost-based pricing, because it is much easier to be understood from a decision-making perspective. Raju and Zhang [18] claimed that this pricing strategy had been adopted by an overwhelming majority of U.S. companies. One of the primary reasons to adopt this strategy is it is concrete and tangible. It can also be considered as "fact"-based pricing. Despite the fact that many pricing experts emphasize value-based pricing [31, 32, 33], cost-based pricing is still common, because it can help decision-makers set a baseline to charge customers for the minimum price so that they can at least cover Capex. Moreover, cost-based pricing can articulate a unit cost and provides a measurement for benchmark comparison. It becomes one of the managerial tools for many decision-makers to drive CSP's business performance. Last, the components of cost are the essential element of Cost and Benefit Analysis (CBA) so that a decision can be made realistically [34] and a cloud price can be validated internally.

In practice [35], value-based pricing is often used far less than the other two pricing strategies, and market-based pricing is the most popular strategy and followed by cost-based pricing,

[14]Intrinsic Value – A value can be isolated by its own.
[15]Extrinsic Value – A value is dependent on others, or instrumental value.

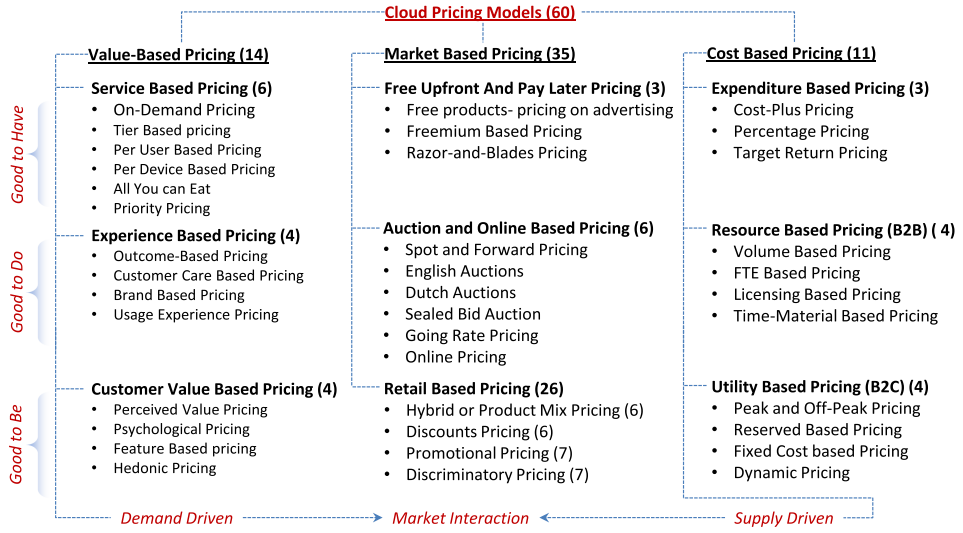

Fig. 5. Taxonomy of overall cloud pricing models.

as shown in Figure 13 (Online Appendix E). This result indicates that value-based pricing is much more challenging to be applied due to a value estimation of cloud customers' experiences, satisfaction, and perception. Nagle et al. [17] proposed a practical solution of value metrics that consists of six value cascades to implement value-based pricing: value creation, value communication, price structure, pricing policy, price setting, and price competition. Based on three pricing strategies, three value propositions, and the combination of 3×3 value metrics, we can create a taxonomy of cloud pricing.

3 TAXONOMY OF PRICING MODELS

According to the above criteria of cloud pricing, we can further map onto 60 different pricing models that are determined by four factors of value, fact, supply, and demand, which underpin a comprehensive framework of taxonomy. It consists of nine different categories that are form 3×3 matrix, as shown in Figure 5 and Figure 6. Each category of pricing has between three and six pricing models except retail-based pricing models. From Section 3.1 to Section 3.9, we will first define each category and then will explain why some models have been adopted by CSPs and others not. Finally, we will link each category to today's cloud pricing practice.

Notice: It is also possible to carve various pricing models at different joints. It may lead to one price model to be mapped onto different categories and different strategies. This is dependent on many factors, such as a business strategy, investment budget, cloud technology, competitive market environment, and targeted customers. Ultimately, it is dependent on a value proposition. In practice, we combine various pricing models to form a new pricing category and to achieve a particular tactical object. This taxonomy, together with three pricing strategies and a 3×3 value matrix, defines a conceptual framework to help cloud decision-makers to examine cloud price models systematically.

3.1 Service-based Pricing

Service-based pricing is to distinguish with physical product-oriented pricing. It emphasizes an intangible part of value delivery. The example of service-based pricing is applied to banking, legal consultant, airline, insurance, travel, hospital, and so on. The aim of service-based pricing is to

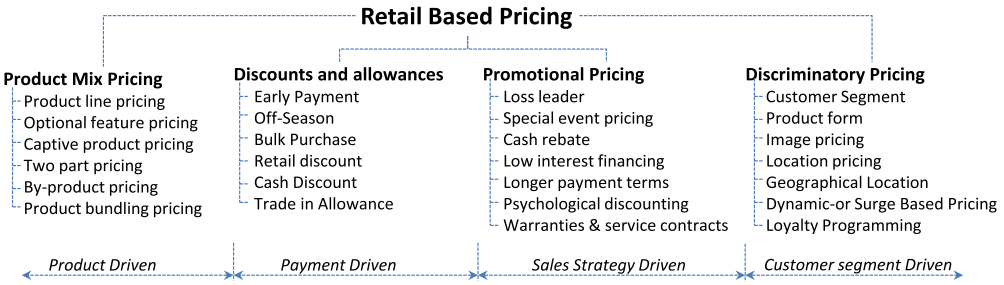

Fig. 6. Taxonomy of retail-based pricing.

focus on the value of "good to have" for cloud customers. In comparison with the other two value-based pricing categories, the value of service-based pricing focuses on value consolidation for cloud services. Many CSPs of SaaS adopt service-based pricing models, such as Salesforce.com and Azure. The value of pricing is measured by the unit of a tier, a level, per device, per user, and a priority. To some extent, it can also be considered incentive-based pricing, because this category is determined by an incremental value of delivery. The advantage of this category of pricing models is their values can be identified and predicted. There are six different models of value pricing: on-demand, tier-based, per-user-based, per device-based, all you can eat, and priority-based pricing. The value measurement of these models may be dependent on a Service Level Agreement (SLA) [63, 91]. Although service-based pricing is closely associated with performance pricing due to SLA measurement, the former focuses on the pricing of service contents while the latter aims at the pricing of the performance required.

The concept of service-based pricing could also be mixed with resource-based pricing, because both categories of pricing may involve some components of intangible inputs and outputs. However, the service-based pricing focuses on value-added service, while resource-based pricing emphasizes the requirement of various inputs. The typical example of service-based pricing for cloud service is "on-demand" or PAYG, which is one of five essential characteristics of cloud service [126].

3.2 Performance-based Pricing

This pricing category means the value of service estimation is based on a set of tangible metrics, which is often measured by service performance metrics, such as the specified reliability of a cloud service or utilization rate of a limited resource (e.g., cloud infrastructure or data center capacity). The aim is to sell the new service values to customers for their performance requirements, such as end-users response time, network throughput, latency, security, and scalability. It may also be considered experience-based pricing. According to M. McNair's definition [40], "performance-based pricing is an arrangement in which the seller is paid based on the actual performance of its product or service."

A typical example of performance-based pricing is online advertising payment, which is dependent on measurement data, such as the number of clicks or purchases [41]. Other applications include telecom services (such as multi-party video conference, mobile apps, satellite connectivity, etc.), which the service prices rely on its specified performance metrics. This pricing category is often connected to the customer's business outcome. The basic idea is to make sure that a CSP's services meet the customer's business objectives or value. The reward of this model is that both parties' values are aligned. By doing so, the CSP will not undercharge the pricing, and a cloud customer will be given the performance guarantees for the services. The advantage of these models

can become "win-win" pricing models and be fair to both parties. From a customer perspective, this model shifts the uncertainty risks to a CSP. However, not every performance metric can be quantified or determined. Sometimes, the performance metric is quite complicated. For example, how do we determine the length of the period for the number of clicks for one online adverting camping? Often, the advertising campaign time may take longer than what was initially expected. In practice, the performance-based pricing models can be subdivided into four different models based on customer's experiences of "good to do." They are outcome-based, customer care-based, brand-based, and usage of experience-based pricing. In comparison with other categories of value-based pricing strategy, the performance-based pricing is tangible because of the definable performance metrics. In a cloud practice, many B2B cloud services emphasize on performance-based pricing, which a CSP offers a guarantee performance, such as five-nine service reliability or 20 Gigabit/s network throughput and in return to charging a premium price.

3.3 Customer Value-based Pricing

Customer value-based pricing is about setting a price from a subjective view of a customer. It focuses on the customer's value delivery. This category of value-based pricing consists of four pricing: namely perceived value, psychological, feature, and hedonic-based pricing. A customer's core value is the main reason to build various price models. If customers believe the cloud service value offered by CSP can be identified for their future values, then they will be willing to pay (W2P) for it. These four models are constructed by the context of perception, psychology, sociology (large environment) and economics (utility). The primary advantage is that it allows a CSP to maximize its business profit and lead the cloud market. The primary challenge is how to define the value metrics by measuring customers' subjectiveness value for "good to be." In comparison with other models, hedonic pricing [113] is a good model to estimate new service values if the historical dataset is available. These models can be effectively applied to an ever-changing environment in term of new cloud features (characteristics). However, not every feature of service would be "good to be" for every customer. As a result, a decision of selecting cloud service features in corresponding to charging price could become a challenge from a CSP perspective.

3.4 Free Upfront and Pay Later Pricing

Due to the market competition, many CSPs adopt a "Free upfront and pay later" pricing model. The idea is to leverage free products with minimum features so that the pricing model can capture more customers and make the profits from premium customers. There are often three types of pricing models: free products-pricing on advertising, freemium, and razor-and-blades pricing. With a free product pricing on ads model, it can stimulate customer's demand, and customers can enjoy free products. The bad news for customers is that they could waste a lot of additional time trying various free products. Moreover, this model requires a sizable market. If the market size is not large enough to offset the cost of free products, then the pricing model is unsustainable. For the freemium model, there are four types of sub-category models: (1) Classic feature-limited freemium (AWS and Dropbox adopt this model). (2) Free trial period (MS Azure and Oracle cloud services). (3) Free software and premium service support (Linux Red Hat), and (4) Unlock the capped speed or bandwidth or unique service feature (mobile apps, gaming and pay-TV services). These models are pricing four different values, which are quantity, a period, quality, and service features. The critical issue is how to draw a line between free and premium services. Recently, AWS began to offer Lambda service or FaaS, which is one of the freemium services in term of quantity (execution times or a number of clicks and memory size/per month).

The razor-and-blades model is similar to freemium, but the main difference is that razor-and-blades emphasizes the concept of regular and consumable components. For example, a provider

may give away or charge a minimum price for the first or not-consumable element, such as a printer, but charge a high premium for a regular and consumable replacement component, such as printer cartridges. The main advantage of this model is it can optimize the product prices and increase sales and maximize the business profits by redefining different values of product components. However, not every product can be divided into "razor" and "blades." Moreover, with the intense market competition, the provider may risk recovering the "razor" cost due to losing returning customers. From a value perspective, these market-based pricing models are "good to have" to consolidate a CSP's market share. Now, many leading CSPs have offered this pricing model for their Function as a Service (FaaS), such as IBM OpenWhisk, GCP, and Azure function services. FaaS-based pricing is one type of free upfront and pay later pricing.

3.5 Auction and Online-based Pricing

3.5.1 Auction-based Pricing. Auction-based pricing is where the auction mechanism decides the pricing. The term mechanism means an established process that is under the control of a designer [129]. Asunción Mochón [46] stated: "Auction is a market mechanism, operating under specific rules, that determines to whom one or more items will be awarded and at what price." The reason for auction-based pricing is that the market price of some products, such as artworks, antiques, and certain rights (radio spectrum licenses), would be best to be settled via pricing bidding mechanisms. Today, numerous products and services are under a hammer from inexpensive items sold on the internet (eBay) to billion-dollar mobile spectrum license. Many commodity products, property, and financial bonds are included. AWS also places its EC2 and S3 under its auction bidding rules.

There are some pros in term auction-based pricing: The speed of the auction is relatively fast. There are no backward and forward processing steps. The price is also very transparent, which the bidder only pays the increment cost at each bid. Moreover, it is fair to all bidders or players who obey the auction rules. The auction process is straightforward and direct. The limitations of the auction are as follows: For a bidder (or customer), they have very little time to think during the bidding process. Subsequently, bid may be much higher than the real value of the goods. Under the auction theory, there are different types of auction mechanisms based on the design criteria. Lawrence M. Ausubel [47] listed about 13 different kinds of auctions: (1) Clock auction, (2) Combinatorial auction or package bidding, (3) Dutch auction (Open Descending), (4) English Auction (Open Ascending), (5) First Price Auction, (6) Second Price Auction, (7) Pay-as-bid action, (8) Revenue Maximization or optimal action, (9) Simultaneous ascending auction, (10) Uniform-price auction, (11) Vickrey auction (Second Price Seal-Bid Auction), (12) Vickrey-Clarke-Grove (VCG) mechanism, and (13) Winner's curse.

In general, auction model category can be classified into four basic types: English Auction, Dutch Auctions, the first price sealed-bid, and the second price sealed bid (or Vickrey) auction [130] based on the bidding rules design. This article only focuses on the auction models that are closely associated with the cloud market. For example, AWS has adopted a modified spot pricing since 2009. The term "spot" literately means the value of cloud resource (Refer to Online Appendix Y for further details of auction-based pricing in term of resource provisioning) at the right moment of settling. It is derived from a commodity market. "Modified" means that AWS spot instance is not a real spot price, because AWS reserves its right to toss or terminate your bided instances at any time by providing 2-minutes warning time in advance if your bidden price is below a fresh bidding price. Currently, only AWS provided the spot instance for public cloud customers. In 2015, AWS offered two modified version of spot instances, namely Spot block and Spot feet, to exploit more customer's surplus value. With the basic spot instance, a customer only bids for one instance. Spot block means the customer can bid for an instance to lock in a finite number of continuous runtime hours (from 1 to 6 hours). For Spot fleet, AWS allows a customer to bid multiple spot instances

from a spot instance pool. AWS also allows customers to mix with different pricing models (e.g., on-demand and spot instance) to form a specified computing capacity, such as 10 VMs that consists of eight on-demand instances and two spot instances. The auction-based pricing model can be considered as designing for a niche and growing market, such as big data analytics workloads. Economically, the aim of spot pricing model is similar to other discount pricing models, such as GCP's preemptible or Azure's low-priority VM, which is to capture more customers' surplus values at a lower end of the demand curve.

3.5.2 Online Pricing. In contrast to offline pricing, the meaning of "online" is the purchasing goods can only be processed via the Internet and cannot be handled offline or in a physical store. However, some online retailers may also offer both online and offline purchasing prices for customers, but the offline price could be higher than the online one. For example, Officeworks provides both online and offline prices, but the offline price is sometimes higher than online.

The upside of online pricing is it can instantly reach a vast number of customers for a provider. The purchase transaction can be made very quickly via an electronic transaction. There are no extra handling expenditures except for some postage costs. It is much convenient for a customer to do online shopping with a comparison of different online prices offered by different online suppliers. Overall, online pricing enables customers to do the shopping and achieve at least six benefits: "shopping at a finger-click," saving time, competitive pricing, a wide range of goods, no time pressure for shopping and reading product information details, and various brands to be selected. The downside of online pricing is high risks of security and privacy issue, lack of or no significant discount, frauds in online pricing, and the extra cost of goods delivered. From a CSP perspective, it can leverage online information via a recommendation system to tail cloud services for a personalized price or price discrimination. As a result, the CSP can improve its both revenue and profit margin.

3.6 Retail-based Pricing

By its name, the retail-based pricing models are based on a small quantity that consumers buy from physical locations or retail outlets (such as discount shops, warehouses, factory outlets, shopping malls, petrol stations, department stores, supermarkets, Sunday markets, etc.). By and large, the retail providers sell products in a small quantity. It is mainly a business-to-customer (B2C) type of pricing model. However, some models are also applied in B2B. There are at least four subcategories of pricing models: product mixing, discounts and allowances, promotional, and discriminatory pricing. Altogether, retail-based pricing has a total number of 26 models. Each pricing subcategory has a different orientation, as shown in Figure 6, which the products nature drives the product mix pricing, the payment option drives the discounts and allowances pricing, the sale strategy drives the promotional pricing, and the customer segment drives discriminatory pricing.

3.6.1 Product Mix Pricing. This pricing subcategory is to mix or combine with different types of pricing models in different ways. Providers can depend on customers' usage patterns to combine different pricing models. The standard practice for cloud services is to combine both on-demand and spot instance pricing models to accommodate both predictable and unpredictable workloads [39]. There are six types of product mixing models, namely product line, optional feature, captive product, two-part tariff, by-product, and product bundling. The primary focus on this subcategory of pricing is the relationship of different products with regard to how to mix various services to achieve the maximum profit by consideration of limited resource capacity, perishable assets, marginal cost, and an optimal mixture of multiple products.

The benefits of these models can boost sales, generate extra revenue or profits, and meet various demands or market segments. However, the main disadvantages of these models are some

customers may feel the frustration of trapping into a cost black hole. Others may decide not to buy at all. It may create a backlash among some premium customers and lead to a bad reputation for service providers. It may also increase the provider's operational costs. The bottom line is how to make a rational decision on pricing that can reflect customers' demands by different market segments. Recently, AWS had implemented this type of pricing model in 2015, which is known as "Spot-Fleet." The distinct advantage is that it can reduce the customer's churning rate and increase sales revenue and profits.

3.6.2 Discounts and Allowances Pricing. Price discounts and allowances are two techniques for a firm to response fluctuation conditions due to market dynamics. The term discount represents a firm to give a pricing reduction because of product promotion, off-season, cash payment, bulk purchase, display, bundle, wholesale, and two-part tariff. This technique is applied to many perishable services. The cloud computer resource is one of these perishable assets. AWS had a few price reductions between 2006 and 2014 [42]. Allowance pricing is another type of price discount, but it is mainly designed for wholesale customers or commercial clients or SME. Overall, this subcategory of pricing models has six kinds of common discount and allowances pricing models, which are early payment, off-season, bulk purchase, retail discount, cash discount, and trade-in allowance.

The goal of this subcategory is "payment-driven" to improve net present value (NPV), which is to increase the return of net cash flow. The benefits of these pricing models are to reduce the stock inventory or to improve the capacity utilization rate, especially for perishable assets, like cloud resources. The main disadvantages of these models are that it may reduce the profit margin and does not have a brand identity. Currently, all three leading CSPs are offering a price discount, such as spot, preemptible, and low priority, for the number of reasons presented in the above Section 2.1.

3.6.3 Promotional Pricing. Promotional pricing is a sales tactic where a discount is given within a specified period. "Most product management teams will create and agree upon a seasonal promotions calendar for their business. The calendar plans out the flow of promotions over a year and is used as a framework that ensures that the available product is sufficient to meet customer demand and maximize business opportunities. Promotions help generate demand and provide for immediate cash flow into a business. Moreover, promotions can help stimulate demand for slow-selling products and so can help reduce product over-stock" [43].

The obvious reward is to increase sales and minimize stock level [44]. The drawback is that it will drag down the overall profit margin. There are seven different pricing models to boost sales, which include loss leader, special event, cash rebate, low-interest financing, longer payment terms, warranties and service contracts, and psychological discounting. The primary focus of this pricing subcategory is sales driven. A typical example is a laptop sale with a cash rebate for a particular model of the laptop. Recently, GCP has started to offer a promotion price for its cloud Tensorflow Process Units (Cloud TPUv2) for US$4.50/per hour [114] in comparison with the standard TPUv3 with a $8.00/per hour. The price is substantially low in comparison with the regular price.

3.6.4 Discriminatory Pricing. Discriminatory pricing means that the pricing model is charging different prices to different customers for the same services. If we look from a value perspective, then it is a customer value-based pricing strategy to charge each customer at the maximum price according to the customer's perceived value, which is the price that a customer is willing to pay. Based on the classification of the microeconomic theory [15], if it is the First-degree price discrimination pricing, then the price is usually dependent on one-to-one negotiation, such as property sale (in private sale). It often requires a lot of effort to capture the customer's maximum value. It is less likely to be applied to a commodity product.

If the discriminatory pricing (or price discount) is dependent on sales volume, then this is called the second-degreediscriminatory pricing. The typical example is a bulk-selling discount in comparison with a single purchase. It is a common practice for wholesale. If the price is based on a specific group of people in society, such as senior citizens and students, then it is third-degree discriminatory pricing. For instance, Microsoft charges student licenses for MS office packages. If we combine different types of discriminatory pricing, then we should have various price models in practice.

Overall, there are seven different types of pricing models: customer segment, product form, image, location, geographical location, dynamic or surge-based, and loyalty programming pricing. The main idea behind this subcategory is customer segmentation, which is to design different pricing models for various groups of customers. Amazon segments its customers by mixing operational revenue streams and offers some advice to business customers [45]. This subcategory of pricing models not only allows a CSP to boost its sales but also to maintain the profit margin. The flip side of these pricing models would increase sales cost, which will ultimately increase the investment risks. The criteria of model classification are two measurements: market segmentation and value principle of "Good to be" to create new values for CSPs. In the cloud industry, the practice of discriminatory pricing is pervasive, especially for cloud storage services. "Bulk-selling or purchase," that is, second-order discriminatory pricing, is a typical example. AWS S3 has a bulk-selling price.

3.7 Expenditure-based Pricing

Expenditure-based pricing means every price model is derivate or built up from the center component—a unit of "cost." In this category, there are three types of pricing models, namely cost-plus, percentage, and target return pricing. The primary driver behind this category is that all price values are proportional to a particular percentage of the total cost.

The benefit of these models is that a CSP knows a targeted return. They are very concise, straightforward, and quick to be constructed. They can guarantee the profit bottom line at least from a modeling perspective. However, these models ignore customer values and market supply and demand. Subsequently, these models may either overestimate or underestimate the market price. Moreover, if the expenditure (cost) item is inaccurate, then this would lead to incorrecnt pricing. Furthermore, the end to end (E2E) or the total expenditure for many large enterprises and government agents are not transparent. Often one cost item will be accounted for multiple times. If so, then this leads to overestimation of the price for offering services. As a result, larger firms or enterprises could lose many business opportunities. However, if some cloud customers have some special requirements, such as regulatory compliance for their running business applications regarding cloud infrastructure, expenditure-based pricing models are good to have. In 2015, AWS released a new pricing model, namely dedicated host to meet customers' compliance requirements. IBM has had a similar price model, which is known as "bare metal," to eliminate the "noisy neighbor" effect. All these models are driven by cloud expenditure (or costs). This kind of pricing model may appear to be contradictory to the characteristics of cloud computing, but it fits into particular business requirements—regulatory compliance, a high degree of security control, streaming applications, and dedicated computing power.

3.8 Resources-based Pricing

Instead of pricing on cost account, resource-based pricing focuses on a consumption base. If may also be considered activity-based pricing (costing) [123]. We classify resource-based pricing as one of the categories of cost-based strategies, because they have some common properties that are associated with the expenditure components. However, not all resource consumption costs money. Some natural resources are free. For example, the natural resource of solar or wind power does

not cost any money. Resource-based pricing emphasizes scalability. Many cloud services are built on resource-based pricing. Chen [124] found that the cloud market or customers have a stronger preference for a particular CSP, or if a CSP can offer higher SLA than its competitors, the the CSP is more likely to adopt resource-based pricing.

Resource-based pricing is common for the services industry. Traditionally, there are many service industries that haveadopted resources-based pricing models, such as e-commerce, airline, travel and leisure, recreation and entertainment, healthcare, and education. Resource-based pricing is also adopted by the IT industry, especially for IT outsourcing purpose. Resource-based pricing aims to offer a better method that allows customers to consume and deploy the scalable resources both efficiently and effectively.

This category of pricing emphasizes resource scarcity [101]. There are four types of resource-based pricing, which are known as Transaction-based, FTE-based, Licensing-based, and Time-Material-based pricing. We can roughly differentiate this category of pricing models by criterion of "good to do." Softlayer and VMware recently launched "VMware virtual data center." It uses resource-based pricing, because it includes all resources of cloud service, even archive storage resource.

3.9 Utility-based Pricing

Nayan B. Ruparelia [39] defined the term of a utility pricing model as follows: "Utility models are metered price models whereby your usage of the service is monitored, and you pay accordingly." His further explanation is that the origin of the model was "from the price plans that utility companies have adopted, they are characterized by regular payments, often monthly, to the cloud service provider."

The term of utility has serval different connotations. (1) From a computer software perspective, it means that the software can perform multiple specified functions. For example, utility software (iOS or Windows) can be utilized to perform the tasks of monitor, mouse, printer, and disk driver. (2) Another meaning utility is very close to the utility function, which is utilization rate for a certain amount of capacity. (3) From a public service perspective, it means an incumbent service provider can provide public services, such as telecom, electricity, gas, water, and public transportation, which are essential to modern society. (4) The economic term of utility is that a person receives satisfaction or pleasure for consumer goods or services. The original meaning of "utility" was coined by Bentham [48], which means the principle of utility or usefulness that is "greatest happiness for the greatest number of people."

For the category of utility-based pricing, the meaning of utility is similar to the metered price for public services. The benefit of utility-based pricing is that every individual can access the cloud service directly via a credit card for the infinite scale of resources without a prerequisite condition, upfront Capex. The flipside is that it is not a good idea to commoditize some new or innovative cloud service features by using this model. Nevertheless, this type of pricing models provides the value of "good to be" for cloud end-user because of OpenStack [125] development. According to various business requirements, usage time, resource commitment, customer segments, and payment types or different workload patterns, utility-based pricing can have different pricing models, namely, Peak and Off-Peak and fixed cost-based pricing. Chen et al. [124] argued that if cloud market demand is less volatile, cloud customers would prefer resource-based pricing. In contrast, if their demand is highly volatile, then they would prefer utility-based pricing.

3.10 Summary of Pricing Models Classification

From both Figure 2 and Table 2 (Online Appendix C), we can see that service-based pricing, especially on-demand, per use-based, and tier-based pricing models, has become a common pricing

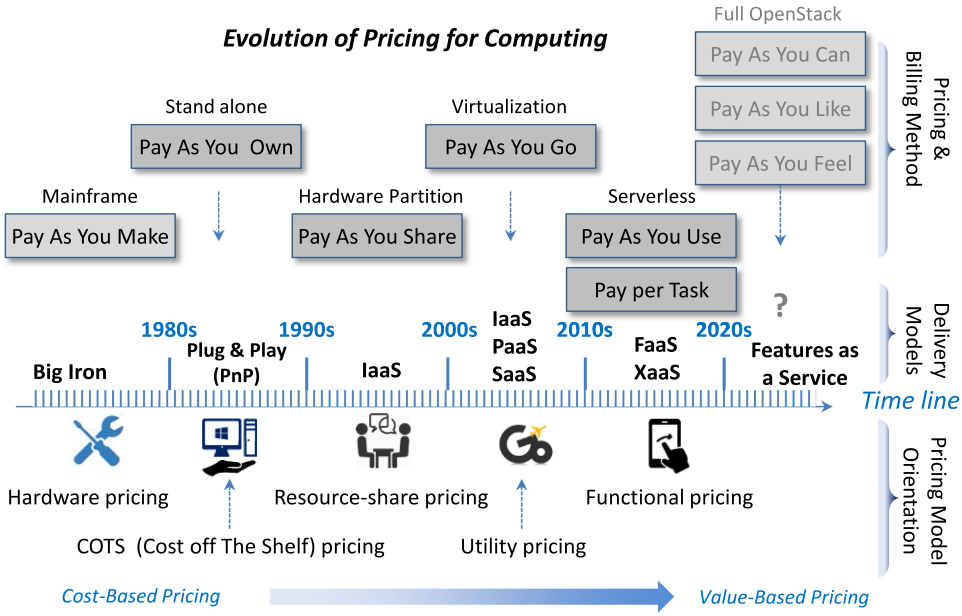

Fig. 7. Evolution of cloud pricing and billing methods.

model widely adopted by many CSPs. The aim of these pricing models is that they can reflect the cloud characteristics of both scalability and "on-demand" or Pay as You Go. If we look back on 40 years of computing history, as shown in Figure 7, then we can see that the billing method is moving from "Pay As You Make" to "Pay As You Use" or "Pay As You Can," the delivery model is moving from "Big Iron" to "FaaS," and the pricing model is moving from hardware based to functional based. Altogether, a pricing strategy is moving from cost based to value based. However, this does not mean that pricing models driven by cost-based strategies will disappear. They could co-exist with various new types of pricing models based on the computing technology adoption lifecycle [4].

As shown in Figure 3, there are approximately seven cloud pricing models or model categories offered by leading CSPs at the moment. From a historical perspective (exhibited in Figure 2), we argue that new pricing models will be created often alongside innovative cloud technologies. We have observed many CSPs, such as Cloudheat [110], Databricks [112], Cisco systems, and RingCentral [111], start to roll out a new pricing model that is supported by a hyper-converged solution to extend cloud computational power to the edge, which is close to the end-user. They call it *distributed* or *fog computing* or *data center in a box*. This solution can eliminate network latency and routing path hops and provide much mobile computation power. Although this type of cloud service may still be in an incubation stage, it could become a significant model. On the other side of the pricing spectrum (Refer to Figure 3), other CSPs, such as Iex.ce [109], Cambridge Intelligence, Arkessa, and Vizolution, extend a cloud resource pool to the global market reach by leveraging blockchains and desktop grid technologies to offer a competitive price, e.g., "Pay-per-Task" (see Figure 7). These practical solutions illustrate that innovative cloud technologies accompanying with new competitive pricing models will stimulate the process of cloud transformation.

There are at least 60 different pricing models for various cloud services. The reason we illustrate 60 pricing models is that different cloud services require different approaches to address various issues of cloud services, such as methods of delivery, payment, promotion, discrimination, and

and so on. The detail of each pricing model is excluded from this article due the limited space. Our analysis results of cloud pricing strategies are similar to Hinterhuber's findings, shown in Figure 13 (Online Appendix E), where the dominant pricing strategy is the market-based pricing strategy (35 pricing models). Overall, we have defined and highlighted many pricing model categories that have been already applied to different industries, especially service industries. Although many of them are not available in today's cloud market, CSPs should not eliminate their imagination to a few pricing models. As Weinman [8] indicated, CSPs should learn from other industries and compete on pricing, not on price alone. Table 4 (Online Appendix F) provides the summary information of these categories of pricing models at a glance.

Throughout the taxonomy of pricing models, we emphasized value-based pricing strategies for cloud services, because the natural characteristic of cloud computing is service. However, it does not mean that cost-based pricing is not important. It often provides a bottom-line price for CSPs. Value-based pricing illustrates the maximum price, which is how much cloud customers are willing to pay, while market-based pricing will give CSPs an estimation of competitive price in the marketplace. If the cost-based pricing can set up the lower bound price, then the value-based price is to estimate the high bound. Market-based pricing gives a price variation between the lower and higher bounded prices. Cloud pricing strategies, tactics, and models are mainly dependent on various cloud services features, cloud technologies, targeted customers, market environment, cloud orchestration, and and so on.

4 SURVEY ON PRICING MODELS

During the past decade or so, hundreds of papers have been published regarding cloud pricing models. Many pricing models can be considered as an extension of grid, cluster, distribution, high performance, parallel, Peer to Peer (P2P), network, and utility computing. Based on the framework of our taxonomy, the following survey will be organized as three cloud pricing strategies. We selected published works between 2019 and investigated with a deep-diving approach. The compelling reason to select these research works is that the majority of studies proposed either new mathematical solutions or novel ideas for the pricing models. The goal of this survey is to transform various mathematic models of cloud pricing into a defined taxonomy in an economic context.

According to a topic of each paper and its contents, we classify References [49, 50, 52, 55, 56, 61, 62] as market-based pricing, References [60, 77, 78, 80, 115–118, 120, 121] as cost-based pricing, and References [84–86, 91, 93, 95, 113] as value-based pricing. We highlight the uniqueness of their ideas, new concepts, and the contributions of each paper. Moreover, we show their relationship, whether it is a continuation of previous work or the original work.

4.1 Pricing Models of Pre-Cloud Computing

In late 1999 and thr early 2000s, Buyya et al. [49, 50] proposed a computational economy framework to regulate grid computing based on market supply and demand. The basic idea is to provide a set of different pricing models that can optimize grid resources and various objective functions through trading and broker services on an open commodity market. The authors introduced at least seven different types of pricing models: commodity market, posted price, bargaining, tendering/contract-net, auction, bid-based proportional resource sharing, community/coalition/bartering, and monopoly and oligopoly models. However, the authors also noticed there were many challenges [51], such as managing grid resources, leveraging grid technologies to allocate grid resource, and implementing different pricing models. As a result, many proposals of pricing need further consolidation.

When virtualization became a mature technology during the 2000s, cloud computing services were on the horizon. Based on many years' research, Buyya et al. [52, 53] argued that the

paradigm had shifted. The authors proposed the architecture solution for market-based pricing for cloud resource allocation. The solution was an extension of grid computing [54], which is to leverage third-party services (or a cloud broker) to allow cloud consumers to utilize global cloud infrastructure effectively. Buyya's pricing model can be considered as broker-driven pricing based on the assumption of a commoditized computing resource. The idea of global cloud or multi-cloud service providers was innovative at that time. It can be implemented by the serverless[16] container[17] technology, which has emerged recently [109]. The aim of Buyya's cloud pricing solution is to increase cloud resource efficiency.

4.2 Market-based Cloud Pricing

Following a similar line of reasoning, Toosi et al. [55] developed a novel algorithm in combination with different cloud pricing models that allow a CSP to optimize its cloud capacity (or cloud infrastructure efficiency) for cloud business revenue maximization. The main contributions of their proposal are as follows: (1) present a stochastic dynamic programming technique to calculate the maximum number of reserved instances that a CSP can offer to cloud customer for its revenue maximization. (2) Due to the computational complexity of dynamic programming technique, the authors provided two heuristic algorithms. (3) The paper created a framework that is validated by large-scale simulation dataset provided by Google. The framework can be presented by the following four equations shown in Figure 14 (Online Appendix G).

Equations (1), (2), and (3) are three constraints. Equation (4) is the sum of quantity multiplied by the price units to achieve three revenue streams based on three price models: reserved, on-demand, and spot (or auction-based pricing). The paper presented a novel idea about how to maximize cloud revenue within the fixed amount of cloud capacity based on three existing AWS cloud pricing models. However, there are some gaps regarding pricing model assumptions: (1) the revenue function excluded the cost component; (2) AWS is charging on hourly base for on-demand instance while Google Cloud Platform (GCP) is charging on a per-minute base; (3) based on the AWS price model, spot instances can be terminated in 2 minutes advance warning time. So the l_t^s can be set to zero at any time, and s_t can also be set to zero if there is an issue of cloud capacity contention. In other words, there should be one more assumption regarding the termination of a spot, because AWS does not charge customers if the spot is less than an hour. The remaining challenge is how to model an arbitrary behavior of the instance termination.

Similarly, Xu et al. [56] tackled the same problem by introducing a dynamic pricing model that can be traced back to Gallego's work [57]. The main idea of their dynamic pricing model was to assume that both the arrival $f(p)$ and departure $g[f(p)] = k[1 - (f(p)]$ (where, $k > 0$) rates for AWS spot instance demand are a Poisson process. If the optimal stochastic policy changes price continuously (or the price change is a continuous variable), then the expected revenue function E_u and maximum profit $J^*(x, t)$ are shown in Figure 15 (Online Appendix H).

The essence of Xu's work in Figure 15 (Online Appendix H) is the derivative equation $\frac{\partial J^*(x,t)}{\partial t}$ for CSP's profit maximization. It equals the optimal spot price multiple with the quantity of spot that is subject to both arrival and departure rates. The main contributions of this article offer an alternative pricing model for CSP to model its spot instance's price dynamically. This means that a CSP reserves its right to change the spot price at any time. The authors argued that this pricing model could provide a control mechanism for CSP to utilize its limit cloud capacity better. However, few assumptions need further consolidation.

[16] Serverless – a cloud computing execution model without a defined server – event driven application deployed model.
[17] A container is a package of software code that is fit together and allow cloud user to run an application quickly and reliably among different computing platforms.

The observation of spot price variation within a narrow band could be validated. The price could be accurate for a particular instance in the past. However, it is quite challenging to be generalized across all instances, zones, and regions in the future environment. Joshua Burgin (one of the general managers from AWS) indicated: "Prices for instances on the spot market are determined by supply and demand. A low price means that there is more capacity in the pool than demand. Consistently lower prices and lower price variance mean that the pool is consistently underutilized. This is often the case for older generations of instances such as m1.small, c1.xlarge, and cc2.8xlarge" [58]. AWS "Spot Bid Advisor" shows many instances are frequently outbid shown in red in comparison to its on-demand price in Figure 16 (Online Appendix I).

In some cases, the spot price reached a ridiculously high price, $999.00 [58], which was well above the on-demand price. This phenomenon indicates that the spot instance price variant with time is neither convex nor continuous. As Gallego [57] stated, "the stochastic optimal policy changes prices continuously and thus may be undesirable in practice." Moreover, both arrival and departure functions are defined as more like a power function rather than a Poisson distribution function, which the paper assumed as follows:

$$f(p) = k(1-p^a)^b,\ g[f(p)] = k[1-f(p)] \quad (\text{where, k} > 0, \text{a} > 1, 0 < \text{b} < 1) \tag{1}$$

In addition, the model also excluded the cost component for CSP's revenue maximization. Their interpretation of Greenberg's [60] works could be inaccurate. The paper also assumed that cloud customers are the price takers, because AWS has full control of the spot instance based on both arrival and departure rates.

So the question is how the AWS controls its spot instance and what is a mechanism behind the AWS' spot instance bidding processing? Before our further investigation of AWS spot instance, it is crucial to understand how the spot bidding process works. AWS bidding mechanism is very similar to the second-price sealed-bid auction (or Vickrey auction). In contrast to the English (open) auction process, it is a blind auction, in which all the bidders submit their bidding prices simultaneously without any pre-knowledge of other bidding prices—the highest bidder wins the cloud instance time slot at that time. However, the price the highest bidder pays is slightly higher than the second-highest bidding price, not the highest bidding price. For example, the reserved price of the highest bidder is $2.00, but the next bidding price is only $1.00, so the highest bidder only pays $1.01, not $2.00.

AWS might have its own reserved price with different types of spot instances across different regions and zones based on the availability of its resource capacity after satisfying its "on-demand" and reserved customers. When a new bidder submits a fresh bidding price that is higher than the old bidder's reserve price at any time, the old bidder has a warning time of 2 minutes to terminate his or her running instance. In this case, AWS will not charge the customer if the instance runtime is less than 1 hour. The existing customers can either revise their upper ceiling reserved price or move their workloads to "on-demand" instance. As we illustrated above, the bidding price might be well above the "on-demand" price. It might sound irrational. However if a customer only pays a very short period, the price will become acceptable because the average spot price is less than "on-demand." As a result, Xu's spot pricing models require further consolidation.

Recently, AWS has capped four times of "on-demand" price as the highest bidding price. Moreover, AWS also offers up to 6 hours spot instance (spot block in 2015) to accommodate different types of workloads. These new combinatorial pricing schemes will change the bidding game. Furthermore, AWS also provides historical spot pricing records and help customers to form their pricing bidding strategy. Based on AWS historical spot pricing dataset, Ben-Yehuda et al. [61] provided a different interpretation of AWS spot pricing, which they argued the AWS spot instance has its reserved price. Their conclusion is based on a reversed engineering and traceable dataset (from Tim

Lossen's Cloud Exchange and Kurt Vanmechelen's Spot Watch) in April 2011. They illustrated that the high bound of a spot price is set to reflect a market-driven (auction-based pricing) mechanism while the lower bound price is reserved within a narrow band, which can be presented as:

$$\delta_i = -a_1\delta_{i-1} + \varepsilon(\sigma), \text{ and } p_i = p_{i-1} + \delta_i \tag{2}$$

where δ_i is the narrow band, a_1 is the coefficient, $\varepsilon(\sigma)$ is the white noise, and p_i is a price at any time "i." It is an empirical observation. The goal of the paper was to help cloud customers to understand AWS spot mechanism to bid the spot price.

Zheng et al. [62] intended to answer a similar question as Ben-Yehuda. They presented spot price bidding strategies for different types of workloads. The authors' conclusions were their bidding strategy could reduce 90% of the cost in comparison with the "on-demand" price. The paper assumed two types of scenarios, which are one-time bidding and continuous bidding strategies. For the one-time bidding strategy, the cloud consumers can achieve the lowest possible bid price p^* illustrated in the following Figure 17 (Online Appendix J).

Zheng's work can be summarized into three main contributions for the AWS spot instance pricing bid strategy: (1) Price orientation bid strategy, (2) SLA priority bid strategy, and (3) MapReduce workload application. Based on the authors' observation, they conjecture that only a few users bid for spot instances due to heavy-tailed spot price distribution. However, the gaps of their pricing models are as follos: (1) the authors assumed that the highest spot bid price should be less than the on-demand price, but in fact, the bidding price could be well above the on-demand price (four times higher than on-demand). (2) The maximum revenue function analysis did not include the marginal cost from a CSP perspective. (3) The authors did not give a further explanation of the capacity utilization function. (4) The assumption of uniform distribution for bid prices appears to be contradicting the later contents of bid prices distribution: Pareto and exponential distribution. (5) The paper intended to isolate the issue of the spot resource from other on-demand and reserved resources, but, in fact, a CSP often has a large resource pool for all price models. (6) The assumption that the workload is i.i.d. needs further clarification.

Overall, the spot or auction-based pricing serves well for interruptible workloads. These jobs have some essential characteristics: (1) Running time for the job is unpredictable, (2) it has many checkpoints, (3) the job can continue to run after any stop point, and (4) it works well for stateless applications or processes (the server does not save the client's data that is generated in one session). Based on the paper's final discussion and conclusion, the spot pricing bid strategies are only applied for interruptible workloads.

Since AWS launched its spot instance in 2009, it has generated enormous interests in the academic world. The amount of published papers [63–70, 72] regarding the AWS spot pricing model is overwhelming. The main reason is that this model can offer up to a 90% price discount in comparison with "on-demand" price. The basic idea of a spot pricing mechanism can be considered as analogous to the energy (electricity) market [71]. Many SLA and cost saving-oriented papers proposed some complicated mathematical formulas based on both historical spot price data and subjective assumptions. However, the reality is that AWS can terminate any spot instance arbitrarily, although it gives 2 minutes of warning time in advance. It is quite challenging to model the AWS termination mechanism.

A SaaS company, MOZ's experience of September 26, 2011 [59], provided a typical example, showing that it would take a higher risk to rely on the spot instance alone for SLA services delivery. Because MOZ was out of the bid,[18] all MOZ[19] services were shut down [73]. It took MOZ 14 days

[18] MOZ reserved bid was ∃2/per instance for more than 3 years.

[19] MOZ provides Search Engine Optimization (SEO) web crawler services to its customers. MOZ charges its customers on monthly subscription fee.

to restore its services fully. MOZ has about 26,474 subscribers plus 5,000 free trial customers. If we assume MOZ's customers pay a premium of $599/ month, then the estimated revenue loss is about $8 million in 14 days if we exclude impacts of potential new subscribers, customer experiences, and the company's brand and reputation. This is why MOZ switched its cloud infrastructure from a public cloud to colocation [74] in 2013.

Usually, the spot instance is not an ideal resource for mission-critical applications, but it could be applied to interruptible workloads. This means that customers should understand their workload first and then determine which type of VM instance is best. Some computation-intensive workloads, such as encoding or decoding, rendering, modeling, or continuous integration, cannot generate checkpoints over its multi-hour running period, so it is not wise to select an auction-based price (spot instance). In comparison with AWS, other leading CSPs, such as GCP and Microsoft Azure, do not offer spot pricing model but provide a fixed discount price with limited service features. Overall, AWS's spot instance, GCP's preemptible, and Azure's low priority offers a cost-saving opportunity if the workload type is applicable.

4.3 Cost-based Cloud Pricing

On the topic of the cost saving, Greenberg et al. [60] proposed the cost-based strategy regarding cloud data centers as early as 2008. It provided a rough estimation of infrastructure cost for cloud services. Some critical assumptions of their estimation were 50,000 physic servers or nodes and 5% of interest rate for capital investment, $3,000 per server, three-year lifecycle time and electricity price $0.07/ KWH. The guideline to build its own cloud data center showed in Table 5 (Online Appendix K).

The authors highlight significant issue across many data centers at that time (before 2008), which has a lower utilization rate (less than 10% on average) of data center resources. They identified some approaches to increase the data center efficiency, such as optimize the data center internal network, design market-based algorithms for data center utilization, and improve inter-connected data center network. We argue the estimated costs for the cloud data center are dependent on each case and the location of a data center. For example, the authors assumed the electricity price is $0.07/ KWH. This price estimation is on the lower end [75]. The average price of electricity power cross developed nation (OECD) is US$0.23 [76]. Even in the U.S., the average price of household electricity is around 0.125, and the industrial price is about $0.10. If we use OECD average price and keep other cost items unchanged, then the proportion of each cost component for the amortized cost will be changed dramatically. The portion of the amortized cost of electricity will be double. Moreover, the paper did not include the data center space cost, which is another essential cost item. It could be up to 15% [75] of the total cost of a typical cloud data center. Nevertheless, the paper made a significant contribution to cloud data center price estimation. They are the pioneer of cost-based pricing for cloud services.

In comparison with Greenberg's approximation estimation, Walker [77, 78] laid out the precise costs of both CPU and storage for Net Present Value (NPV) in comparison with AWS EC2 and S3 (or public cloud) presented in Figure 18 (Online Appendix M). According to Walker's calculations with assumptions of 90% of server utilization rate, 5% of a capital cost, and clusters of 60,000 CPU cores capacity, the author concluded that a three-year investment commitment is the optimal term length for purchase case because of the lowest cost per CPU hour. Second, the operational lifespan should be within 10 years. Moreover, if the lifespan is less than two years, then it would be cheaper to lease computational capacity (off-premise). Finally, if the capacity utilization rate is less than 40%, then it would always be more reasonable to use cloud resources (off-premise). Based on the same principle (NPV), Walker demonstrated formula for the enterprise storage cost in the

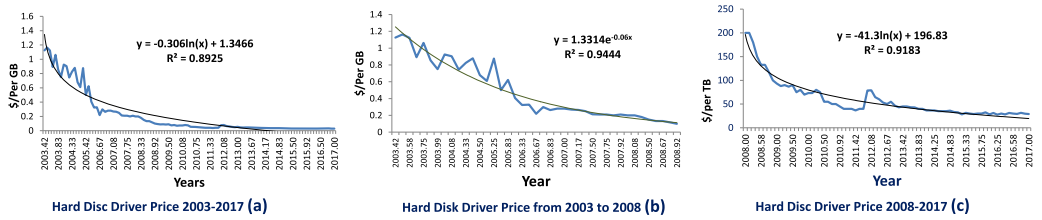

Fig. 8. Hard disk drive price within various years' spans.

comparison between own build (on-premise) or purchases and public cloud (off-premise) shown in Figure 18 (Online Appendix M).

The assumption for cloud storage pricing was based on the threshold levels of storage illustrated in Table 6 (Online Appendix L). This means that CSP often gives a volume discount, which is a kind of linear discount rate.

However, the issue is that the storage price is quite challenging to be generalized, because each CSP will have a different cost-based pricing model for cloud storage (as shown in Table 7 (Online Appendix O)). The price range could be as high as 21 times difference, which is dependent on many storage performance factors. Moreover, each CSP may give different depreciation rates of cloud storage price each year. This means the L_T (Expected annual per GB lease payment) is a time variable, not a constant.

Walker's suggestion was if a decision-maker wants to have cloud storage resource for more than 4 years, the solution of building own storage infrastructure (on-premise) is a preferred option otherwise cloud solution (off-premise) would become a favorite option because of a higher NPV value. The main contribution of Walker's papers was the author demonstrated how to use the NPV to construct a cloud cost-based model by taking consideration of Moore's law or IT assets depreciation within a specified period. However, the predicted cost per GB is dependent on the observation of previous years. Various sources of price data collection could lead to different results. For example, if we adopt McCallum's dataset [79], then the $G_x = 1.3314e^{-0.06T}$ (the depration rate of $ per GB) between April 2003 and September 2008 (see Figure 8(b)) Moreover, if we take the period from 2003 to 2017, then the best format to fit historical HDD price dataset would be logarithmic rather than exponential (see Figure 8(a)) $G_x = -0.306 \ln(T) + 1.3466$. The R-square value is 0.8925. Finally, if we take the time span from 2008 to 2017 and change the price scale from dollar /GB to dollar /TB, then the coefficient of the fit equation would change again: $G_x = -41.3 \ln(T) + 196.83$. The R-square value is 0.9183 (see Figure 8(c)).

This indicates that E_T (a capital cost in year T in Online Appendix N, Figure 19) is dependent on the number of observation years (or data points) and the unit of time span and unit price/per HDD. If these variables are changed, then the fit-equation and its coefficients will also be changed. Subsequently, the decision model is oscillating according to different time spans. Walker's cost-based pricing can be considered as a root of resource performance driven by cloud customer's NPV. If we shift our focus from cloud customer to CSP, then a value proposition becomes an issue on how to optimize the finite capacity of the cloud resource.

Xu et al. [80] proposed a preliminary price model for cloud resources. The basic idea of their model is derived from one of the customer's utility functions: Isoelastic or constant elasticity functions (a particular case = constant relative risk aversion (CRRA) of Hyperbolic Absolute Risk Aversion (HARA)) based on economic utility theory [81]. They argued if a CSP seeks to maximize its revenue and cloud consumers will make rational choices with risk aversion preference, the CSP can have five different strategic options for cloud pricing, namely (1) basic, (2) the firs-order

price discrimination, (3) throttling, (4) SLA performance, and (5) profit maximization, which is illustrated in both Figure 20 (Online Appendix P) and Figure 21 (Online Appendix Q). In Figure 21 (Online Appendix Q), Equations (6) and (7) show how to find the optimal price to maximize CSP's reverence with a limited cloud capacity.

The main contribution of Xu's paper is that it articulated various CSP's pricing choices by exploring the iso-elasticity function as a cloud customer's utility. The author demonstrated that CSP could leverage customers' surplus values $S_v(p)$ to maximize its revenue if there is only one type of utility. However, there are a few practical issues: (1) the customer utility function and alpha-fair utility are two different concepts. One is the utilization rate of the limited amount of cloud capacity, and others have an economic connotation, which is to measure customer's subjective experiences or satisfaction. (2) As authors indicated in the paper, it is challenging to charge cloud consumers with the first-order degree price discrimination because of the price transparency. In practice, it is more likely to adopt the second-degree (volume discount) and the third-degree (different prices to different consumer group) pricing discrimination. (3) The assumption of throttling requires further consolidation, because the characteristics of online pricing, CSP has to declare its performance of cloud resource upfront. If a CSP reduces the specified VM performance (such as CPU speed, RAM, and storage size), then this means it cannot fulfill its service obligation. An alternative option is to declare the cloud performance in a rough range. For example, AWS specify its network performance as low, low to moderate, moderate, high. AWS does not provide a quantitive specification. (4) It is unrealistic to assume that all cloud consumers have the same utility functions. (5) A probability density function $f(v)$ needs further clarification. In addition, GCP and AWS pricing models have different billing units of a VM (see Table 2 Online Appendix C).

Furthermore, the assumption of elasticity $E_d = \frac{1}{\alpha} = 3$ needs a further explanation, because this parameter will impact on the shape of the utility function, which ultimately will determine the optimal price. Subsequently, the level of utility $v = p\sqrt[3]{x}$ If we use the paper's price assumption, p = 0.08 per hour for a small Linux instance, then utility level $v = 0.08\sqrt[3]{x}$. And then the paper used Google, RICC and ANL cluster trace information to validate the utility density distribution. Based on the Alam et al. [82] research work, the workload pattern of Google cluster trace is more like the trimodal pattern rather than a convex. In addition, RICC is a parallel computing cluster 0, and ANL is a grid computing cluster [84]. It would be very challenging for the authors to adopt these datasets for pricing model validation, because AWS is under a commercial cloud environment.

Although the paper had included a cost component in the equation of profit maximization, it only considered the energy cost and excluded other operational expenditure (Opex) and Capital expenditure (Capex) items. Practically, the revenue maximization is not equal to profit maximization. Sometimes, it might mean losing money if the total cost exceeds the sales price, which the higher revenue, the larger deficit is. According to Belleflamme and Pietz [85], the above revenue maximization function (monopoly pricing formula) should be altered as follows:

$$\max_{D_v(p)} \pi(D_v(p)) = D_v(p)p(D_v(p)) - C(D_v(p)) \tag{3}$$

where $C(\cdot)$ is an average cost and both price p and cost $C(\cdot)$ are the functions of demand: D_v, and demand is a function of p. Conversely, the price is also a function of demand: $p = D_v^{-1}(p)$. It would be a challenge to find the optimal value of p.

If we trace a root of Xu's research work, then we can find Xu's pricing model can be considered an extension of Joe-Wong and Sen's [115, 116] research work. The difference was that Xu introduced a probability density function for cloud market demand. Joe-Wong and Sen proposed an analytical or mathematical framework of cloud pricing to optimize resource allocation, fairness, and revenue with a finite capacity of cloud resource. The core idea of their pricing model can be

further traced back to Chiang et al.'s [117, 118] study of network utility maximization (NUM). The essence of Joe-Wong's work can be summarized in the following mathematical pricing models shown in Figure 22 (Online Appendix R).

As authors have noticed that "the function of π_b is a non-differentiable function of the amount of each resource i (e.g. b_i)." Subsequently, the value of b_i is a constant. This result actually reflects on a common practice in the cloud industry that was summarized by Kilcioglu and Rao's work [119], which any price of AWS MV can be presented as a proportion to the price of a base unit of VM configuration. Mathematically, Equation (4) shows this relationship. In other words, b_i is equal to 2^{k-1} for the majority of AWS VM types.

$$p_k = 2^{k-1} p_0 \tag{4}$$

where p_0 is the price of the smallest VM size, p_k is the k size of VM and $k = 1, 2, \cdots$ is the number of VM sizes offered by a CSP. The clear advantage of adopting this price model is that the CSP can build a large VM resource pool at the finest granular level of scalability and minimize a footprint of cloud infrastructure in a cloud data center.

By following a similar principle of the network-oriented root of cloud pricing theory, Shahrad [120] proposed a novel idea of pricing so-called Graceful Degradation (GD) to increase its cloud business profit by improving its cloud infrastructure (data center capacity) utilization rate and efficiency. The key idea of GD pricing model is a self-capping mechanism, which is to "absorb demand fluctuation and reduce spare capacity." In other words, the GD price model is a cloud capacity regulator to smooth Service Providers' (SP, or cloud business customers) demand between peak and valley. Their pricing model was built upon a function that is similar to the Cobb-Douglas utility function (Equation (2) in Figure 23 (Online Appendix S)) for an SP revenue function, which is equivalent to an alpha-fair function (Equation (3) of Figure 23 (Online Appendix S)) regarding the total deliverable capacity and service degradation factor.

The significant contribution of Shahrad et al. work was the novel idea of leveraging fine-grain pricing model to regulate a CSP's limited cloud capacity. It is a hybrid pricing solution to balance customers' demand and limited cloud capacity by brownout mechanisms (similarly to electricity supply). The aim of this pricing model is to find a win-win solution for both customers (gain price discount) and CSP (improve cloud infrastructure utilization rate). Later, Shahrad et al. [121] applied the same concept to SLA delivery. In comparison with many previous works, they included a cost component in a profit maximization function shown in Equation (1) of Figure 23 (Online Appendix S). To achieve the optimal value of c_b(reserved capacity), the profit function $E(p)$ has to be differentiable.

4.4 Value-based Cloud Pricing Strategy

On the topic of value-based cloud pricing, one of the scientific approaches is a so-called hedonic model. It has been widely applied to the consumer price index (CPI) by many OECD countries, such as the U.S. Bureau of Labor Statistics (BLS), Australia Bureau of Statistics (ABS), British Office for National Statistics (ONS), Germany Federal Statistical Office (Destatis), and so on.

El Kihal et al. [84], Weinman [8], Mitropoulou [85] and Zhang [86] proposed various hedonic pricing models for cloud services. El Kihal showed the comparison results among major CSPs (AWS, IBM Cloud, Microsoft Azure, Terremark, and Google App Engine) in term of three cloud characteristics: memory ($ per GB), CPU ($ per CPU) and Storage ($ per 100GB). Their hedonic function can be presented in (Figure 24 (Online Appendix T)). However, their work offered limited information on how the dataset was collected and how many cloud instances were used.

The experiment result indicated that an adjusted R-squared value of the linear regression was between 0.43 and 0.69 (or 0.76 for Terremark). The interpretation of their experiment result is

unclear. Ideally, the constant coefficient of linear regression should be equal to zero, because none would like to pay the monthly fee for no hedonic characteristics (RAM = 0, CPU = 0, and Storage = 0). If the constant is not equal to zero, then it often means a fixed effect. Otherwise, the linear regression model needs further consolidation. Checking the adjusted R square values, it only explained 43% ~ 69% of the data. Both IBM and Microsoft's adjusted R square values were less than or equal to 50%. It might indicate the linear equation is not "goodness of fit."

In comparison to the El Kihal et al. [84] paper, Mitropoulou et al. [85] made some progress of the hedonic method. They explained how and where the dataset was collected, but the author did not generalize the hedonic linear equation. Moreover, the adjusted R2 value of the experiment is only 57.5% and 53.7% for linear and exponential models, respectively. It means the linear model can only explain 1,577 of the total of 2,742 data points. Nevertheless, the paper added three more cloud characteristics (RAM, CPU, Storage, OS, Transfer-Out and Subscription) for the hedonic calculation. The paper's goal was to measure a hedonic price index rather than a hedonic price. But they did not provide a base period to establish a hedonic index for cloud services.

This issue was solved by the work of Zhang et al. [86]. Based on Pakes [87]'s seminal work, the author explained the fundament concept of the hedonic method. The main contribution of the paper was to introduce the time dummy variable for the hedonic model of cloud price to analyze AWS' cross-sectional data between 2009 and 2015 (see Figure 25 (Online Appendix U)).

Based on the experiment results, the adjusted R2 value was 0.9792 for 277 data points. In comparison with previous works, their study made a significant improvement. However, the authors could not collect enough data points for earlier years of AWS cloud services. It might explain that the authors did not provide the coefficient results for time dummy variables. The calculated result of time dummy coefficient had a big issue. Furthermore, the p-value of storage is less significant than other cloud service characteristics. The value of the storage coefficient showed as negative. As the authors concluded, the major issues of the paper are (1) a small sample of data is not enough to lead a reasonable conclusion, and (2) some hidden cloud characteristics were left out.

All the above issues have been solved by Wu et al. [113]. They developed a much-sophisticated hedonic pricing model for cloud services. The model categorized hedonic values with three types of cloud characteristics or three variables, namely intrinsic, extrinsic and time dummy (see Figure 26 (Online Appendix V)). It improved the accuracy of the future cloud price. The significant contribution of their work is to unveil a depreciation rate of cloud service, which is equal to −20%. This rate is equivalent to Moore's law for computer hardware.

In addition to the hedonic method, there are also many other value-based pricing models. For example, Jain's [88] social welfare pricing model focuses on the sum of cloud consumers' value. Performance-based pricing model [89] is associated with cloud resource and applications risks. Feature-based pricing [90] that is related to prioritizing cloud features. Service-based pricing model [91] correlates to the Service Level Agreement (SLA).

Jain's model is much similar to an auction-based spot pricing model. In other words, cloud users can submit their ceiling bid prices (willingness to pay) and CSP can adopt different algorithms to schedule and allocate cloud resources based on the optimized metrics (such as profits, cloud capacity, performance, time of a day, energy consumption, etc.). However, it is quite challenging to be implemented, because it left out the cost components of the cloud services. In general, all customers would like to have free or near-free cloud resource, but "cloud computing will never be free" [92].

Lucanin's [89] performance-based price is mainly driven by CPU's energy consumption costs, namely electricity price, and CPU's temperature traces. The paper claimed that its model could save up to 32% of the cost under certain assumptions. The pricing model is dependent on the workload characteristics and determined by the desired performance of the customers. Overall, the cloud

price is not only dependent on the CPU but also memory, storage size, access bandwidth, and other service characteristics.

Kar's [90] feature prioritized pricing model is to estimate the potential value of the workload to the individual user for a particular context. The paper proposed an integrated approach to price IaaS resource from multi-users perspective. It means the model will aggregate all potential values for all cloud features. The gap is how to define the benefits of these cloud features from various customers, because these values are highly subjective.

Wu et al.'s [91] SLA-based model is a resource allocation or scheduling for SaaS delivery. Similarly to the feature-based concept, SLA can be interpreted as different cloud features, which include response time, provisioning time, data transferring speed, and so on. However, SLA does not only include response time and data transmission speed but also include security, cloud regions, and zone diversity, API compatibility, auto-scaling, vertical and horizontal scaling without a reboot, burstable CPU, backup-snap, 24×7, and so on. Many of these features are quite challenging to be measured or quantified. These service features are included in the cloud service as a whole for a CPS to differentiate its service from other competitors.

While many researchers proposed various value-based pricing models, in theory, AWS first launched the innovative value-based pricing model in 2014, which is known as the Lambda function. It is delivered by the serverless sandbox technology, which is also known as Function as a Service (FaaS). It is supported by Docker[20] container and API technologies. A Docker is the default container runtime engine, and a container can be easily destroyed, stopped and built with minimum effort of setting-up and configuration, which is like an "ephemeral" sandbox.

4.5 Function-based Pricing—Function as a Services (FaaS)

Eivy [93] argued that the serverless sandbox allows cloud customers to have infinite cloud resources with flexibility of vendor-free. In short, if all CSPs support Open API, then cloud users can quickly switch among the different CSPs without worrying about any software compatibility issue (vendor-locked in). The price of AWS Lambda function consists of two components that consist of Hit Pricing and Compute Pricing (Memory allocation). AWS [94] and Sbarski [95] showed the details of how to calculate the total cost of AWS Lambda function. We can use Equation (5) to calculate the AWS Lambda price.

$$P_t = h_r + m_r = (\alpha \lceil X_{100} \rceil h - k) \times r_h + \left(\frac{\alpha \lceil X_{100} \rceil h}{10} \frac{R}{y} - g \right) \times r_m \tag{5}$$

where P_t is the total price of the Lambda function for a monthly bill, $h_r$is the hit price, and m_r is a memory resource price. a is the constant value of second per month = 2,628,000. $\lceil X_{100} \rceil$ is a ceiling function for the roundup integer of code execution time/per 100ms. h is the hit rate/per second (execution rate of computing request). If the user's code execution time is less than 100ms, for example, X_{100} = 85ms, then the ceiling function "$\lceil X_{100} \rceil$" is equal to (or normalized) to 1, if the code execution time is more than 100 (e.g., X_{100} = 101), "$\lceil X_{100} \rceil$ is equal to 2. R is the allocated memory resource, e.g., 256 MB-second. y is the baseline memory 1024MB-second (reference price). $r_h$is the price rate \$.020/per million hits (Lambda@Edge. r_h = \$0.6). r_m is the price rate = \$1.667E-06/per 100ms for 1024Mb-s. k is the free allowance of the first one million hits/per month (If the hit rate is less than about 23 hits/per minute, then it would be free for compute resource. However, Lambda@Edge has no free allowance). g is the free allowance of 1024Mb-s is 400,000 GB-second/per month. For instance, if a cloud user has an application code that has 50 hit/per

[20]Docker – a platform to pack an application with all the dependencies objects into a single standard unit for the deployment.

second and code execution time is 125ms, and the memory size is allocated to 256MB/per 100ms, we should have $h = 50$, $\lceil X_{100} \rceil = 2$, y = 1024MB/per 100ms, The total monthly bill is

$$P_t = h_r + m_r = (\alpha \lceil X_{100} \rceil h - k) \times r_h + \left(\frac{\alpha \lceil X_{100} \rceil h}{10} \frac{R}{y} - g \right) \times r_m = (2{,}628{,}000 \times \lceil 125_{100} \rceil \times 50 - 1{,}000{,}000) \times 0.0000002 + \left(\frac{2{,}628{,}000 \times \lceil 125_{100} \rceil \times 50}{10} \frac{256}{1024} - 400{,}000 \right) \times 0.000001667 = \$52.36 + \$10.95 = \$63.31/\text{per month}.$$

However, if the execution time of code can be reduced to less than 99ms, then the monthly bill can drop down $31.56/per month. From a CSP perspective, this pricing model allows CSP to allocate 75ms (200ms–125ms) compute execution time for another user. From the cloud consumers' perspective, they only pay what the code execution time slot, which is called as "Pay As You Use" (PAYU) or Pay per Task (P/T). (Remark: the Lambda function price excludes the cost of storage, API gateway, and data egress.)

This pricing model might indicate the trend of cloud pricing model is moving towards a much more flexible direction, and billing method becomes PAYU and P/T rather than upfront lump sum payment. But, the disadvantage of this model is if the number of hits/per second is remarkably higher, the cost of Code of Demand (CoD) could be out of control. Sometimes, it could be three times higher than the on-demand price [93]. Overall, the new pricing model is to support FaaS that is working with a new platform or orchestration, such as AWS' Cloud Watch, Rackspace's OpenStack, and Google's Kubernets. Following AWS' lead, Google Cloud Platform (GCP) and Microsoft Azure also launched Functions as a Service (FaaS) platform in early 2016. IBM started to offer OpenWhisk in 2016. All CSPs have a very similar pricing model for serverless computing (see Table 8 (Online Appendix W)).

4.6 Summary

We reviewed the number of papers regarding cloud pricing models from 2008 to the present. Among them, we carefully selected 13 papers and presented a deep-diving analysis of these research works, which can be summarized in Table 9 (Online Appendix X) according to the framework of three pricing strategies defined by a value proposition.

From Table 9 (Online Appendix X), we can conclude that the primary purpose of cloud pricing models is to maximize business revenue, to improve cloud resources efficiency, and to minimize cloud infrastructure costs. The common trait of early pricing models was oriented by the cost-based pricing in research. Many studies mainly focused on the utilization of the cloud infrastructure. Walker's two papers, Greenberg's and Joe-Wong's studies provided a good example. When cloud computing has become the mainstream computational resource, especially after AWS launched its auction-based spot-instance in 2009, the research focus had been shifted to market-based pricing. Xu, Ben-Yehuda, and Toosi proposed their pricing solution that includes on-demand, reserved and spot-instance models for CSP's revenue maximization. They emphasized on how to balance the limited cloud resources with various market demands. Just recently, El Kihal, Mitropoulou, Zhang, and Wu proposed the hedonic method to evaluate CSP's pricing for new cloud service features, which is to consider cloud pricing from a value-based perspective.

The differences of three pricing strategies are that value-based pricing is driven from the demand side while cost-based pricing is oriented by the supply side and the market-based pricing is to focus on the equilibrium of supply and demand. The primary goal of having different pricing strategies and generating multiple price models is to capture more surplus value under a cloud customer's demand curve. If we use ASW as an example shown in Figure 9, then we can see that the current AWS' pricing strategy can capture more customer surplus values (diagram B) than the one of 2009 (diagram A), because innovative service features alongside with new price models have been

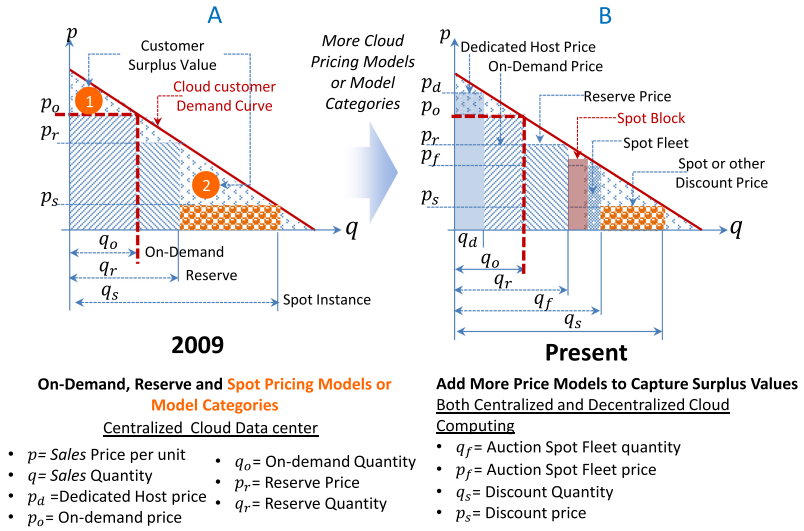

Fig. 9. More pricing models to capture customer's surplus value.

created. It implies that the new cloud price models underpinned by new cloud technologies can maximize cloud business profit for an innovative CSP.

5 CONCLUSIONS AND FUTURE DIRECTIONS

This survey clearly illustrates the point of "the farther backward you can look, the farther forward you are likely to see" [128] for cloud pricing modeling. Based on the review of taxonomy and survey study, we can conclude that cloud pricing is moving further away from a physical box-oriented model to an abstract sandbox-based model. Many CSPs start to offer cloud pricing based on an abstract layer of cloud resource (see to Figure 10). To some extent, pricing of the serverless sandbox can be considered as modeling No Operation Systems[21] (No OS or NoOps), which is an evolutionary direction from an isolated development environment to an integrated environment of both development and operation or DevOps.[22]

However, it does not mean that cloud users can ignore the underlying cloud infrastructure, such as cloud security, workload balancing, horizontal or vertical scaling, auto-failover or high availability, and disaster recovery. All these cloud features will be a part of CSP's responsibility. They become a part of performance measurements or service-based pricing. Cloud customers do not have to get their hand dirty to tune these cloud features directly. They only need to adopt and monitor them and make sure they can be delivered. This is why Kubernetes, Apache Mesos, OpenStack, and Docker Swarm has emerged as an essential tool for cloud transformation.

As a result of the current cloud transformation, we observe that each leading CSP often leverages its business and technology strengths to offer its unique cloud services with innovative pricing models. Based on cloud service delivery models, we argue that AWS can be considered as online retail-oriented pricing for its IaaS delivery, which "AWS brought the Amazon experience to computing resource delivery" [125]. Azure is software application-oriented pricing for its SaaS

[21]NoOps – A programming development approach that allows developers to focus on application development and leave activities of interactions with operation system administrations to a software automation process. It means to take advantages of Platform as a Service (PaaS) to automate application deployment process.
[22]DevOps – it means an integrated process to streamline software planning, building, programming, testing, releasing, deploying, operating, monitoring, and lifecycle.

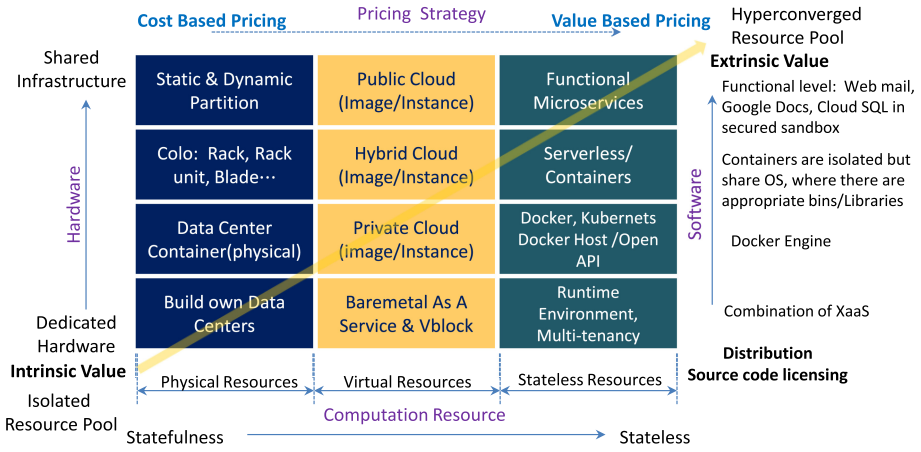

Fig. 10. Future trends in cloud technologies and cloud pricing strategy.

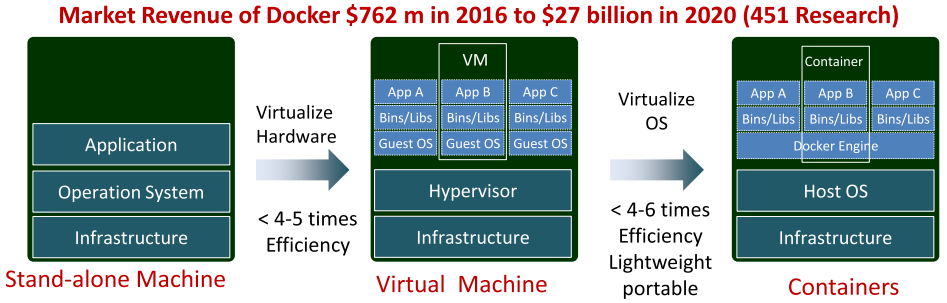

Fig. 11. Estimated market revenue of docker and improvement of computational resource efficiency.

delivery. GCP is search engine optimization (SEO) oriented pricing for its PasS delivery. The other CSPs can leverage their own cloud expertise and strengths for XaaS delivery, such as e-healthcare, cyber-security, Supervisory Control and Data Acquisition (SCADA), Internet of Things (IoTs), and Business Intelligence Analytics.

Overall, the cloud computing technologies and cloud pricing have four possible development trends, which computational resource has moved from statefulness to stateless, IT infrastructure has been transferred from dedicated to the shared base, software development has been gradually shifted from mutability to immutability, and cloud pricing is moving from cost-based to value-based pricing strategy (shown in Figure 10). These trends are leading towards a hyper-converged resource pool for cloud services delivered. We can further elaborate on these trends in the context of hardware, software, and resource architecture (see to Online Appendix Z).

All these cloud developments do not only emphasize the value of hardware but also underscore the value of running business application. 451 Research estimated that the market revenue of Docker would grow more than 35 folds from $761m in 2016 to $27billion in 2020 [96]. The fundamental reason behind this market growth is the efficiency improvement of cloud resource. The initial phase of a cloud transformation from physical to virtual can improve about 4 to 5 times efficiency by reducing cloud infrastructure footprint and cloud data center idle time. The container-based clouds can further increase efficiency about 4 to 6 times shown in Figure 11.

Figure 11 indicates that serverless, Docker container, Open API, DevOps, Desktop Grid, and Microservices will underpin new cloud pricing models. From a CSP's perspective, the implication of the new cloud technologies allows CSPs to meet a challenge of demand fluctuation and maximize its revenue and profit with a finite amount of cloud resources. From a cloud consumer's perspective, it means flexibility of vendor-free, scalability, and Opex minimization. On the basis of these evolutionary trends, we can identify four potential challenges of future cloud price modeling:

- How to move from pure cost-based to both value-based and cost-based pricing.
- How to drive from statefulness to stateless resource pricing.
- How to transfer from mutable to immutable pricing.
- How to price cloud services for both intrinsic and extrinsic features by consideration of cloud infrastructure lifecycle.

Nagle's seminal book [17] provides some clues to deal with these challenges. One of the proposals is to establish or consolidate a value-based metrics from a customer's perspective, which is to create proactive pricing strategy to understand how, and when to satisfy customers' application and meet all their expectations while a CSP can maximize its cloud profit.

REFERENCES

[1] Louis Columbus. 2018. Roundup Of Cloud Computing Forecasts And Market Estimates. Retrieved from https://www.forbes.com/sites/louiscolumbus/2018/09/23/roundup-of-cloud-computing-forecasts-and-market-estimates-2018/#5f62d5b2507b.

[2] STAMFORD Conn. 2018. Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17.3 Percent in 2019. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019.

[3] Peter Burris, Dr. Ralph Finos, David Floyer, and Stu Miniman. Wikibon's 2018 True Private Cloud Forecast and Market Shares. Retrieved from https://wikibon.com/wikibon-2018-true-private-cloud-forecast-market-shares/.

[4] Geoffrey A. Moore and R. McKenna. 1999. *Crossing the Chasm.* HarperBusiness, 5–7.

[5] Vmware. 2019. VMware TCO Comparison Calculator. Retrieved from https://tco.vmware.com/tcocalculator/.

[6] AWS. AWS Total Cost of Ownership (TCO) Calculator, 2008-2019. Retrieved from https://awstcocalculator.com/.

[7] Laura Shiff. 2018. Gartner Magic Quadrant for Cloud Infrastructure as a Service. Retrieved from https://www.bmc.com/blogs/gartner-magic-quadrant-cloud-iaas/.

[8] Joe Weinman. 2012. *Cloudonomics: The Business Value of Cloud Computing.* John Wiley & Sons, Hoboken, NJ, p. 160.

[9] Benedikt Martens, Marc Walterbusch, and Frank Teuteberg. 2012. Costing of cloud computing services: A total cost of ownership approach, In *Proceedings of the 2012 45th Hawaii International Conference on System Science (HICSS'12).* IEEE.

[10] Rajkumar Buyya, James Broberg, and Andrzej M. Goscinski. 2011. *Cloud Computing: Principles and Paradigms.* Wiley, Hoboken, NJ, 3–37.

[11] Jatinder Singh and Vikas Kumar. 2016. Multi-disciplinary research issues in cloud computing. *J. Inf. Technol. Res.* 7, 3 (2014), 32–53.

[12] George Pallis. 2010. Cloud computing: The new frontier of internet computing. *IEEE Internet Comput.* 14, 5 (2010), 70–73.

[13] Morris Engelson. 1995. Pricing strategy: An interdisciplinary approach. *Joint Management Strategy*, 6–17.

[14] Paul Belleflamme and Martin Peitz. 2015. *Industrial Organization: Markets and Strategies.* Cambridge University Press, New York, 27.

[15] Hal R. Varian. 1989. *Price Discrimination Handbook of Industrial Organization 1.* Elsevier, Amsterdam, 597–654.

[16] Irvin B. Tucker. 2017. *Microeconomics for Today* (9th ed.) Cengage Learning, 57–90.

[17] Thomas T. Nagle, John Hogan, and Joseph Zale. 2011. *The Strategy and Tactics of Pricing: A Guide to Growing More Profitably* (5th ed.). Pearson, Boston, MA, 72–93.

[18] Jagmohan Raju and Z. John Zhang. 2010. *Smart Pricing, How Google, Priceline, and Leading Business Use Pricing Innovation for Profitability.* Pearson Education, Inc., 20.

[19] Karl R. Popper. 2012. *The Open Society and Its Enemies* (5 ed.). Routledge, UK, 68.

[20] George Edward Moore. 2004. *Principia Ethica.* Dover, New York, 59–108.

[21] Marc Benioff and Carlye Adler. 2009. *Behind the Cloud: The Untold Story of How Salesforce.com Went from Idea to Billion-Dollar Company—And Revolutionized an Industry.* Jossey-Bass, San Francisco, CA, 103–105.

[22] Alexandru Iosup and Dick Epema. 2011. Grid computing workloads. *IEEE Internet Comput.* 15, 2 (2011), 19–26.
[23] Valarie A. Zeithaml. 1998. Consumer perceptions of price-quality and value: A means-end model and synthesis of evidence. *J. Market.* 52, 3 (1998), 2–22.
[24] William D. Ross and Philip Stratton-Lake. 2002. *The Right and the Good.* Clarendon Press, Oxford, 65–74.
[25] Jagdish N. Sheth, Bruce I. Newman, and Barbara L. Gross. 1991. Why we buy what we buy: A theory of consumption values, J. *Bus. Res.* 22, 2 (1991), 159–170.
[26] Wolfgang Ulaga Samir Chacour. 2001 Measuring customer-perceived value in business markets: A prerequisite for marketing strategy development and implementation. *Industr. Market. Manage.* 30, 6 (2001), 525–540.
[27] Andreas Eggert and Wolfgang Ulaga. 2002. Customer perceived value: A substitute for satisfaction in business markets? *J. Bus. Industr. Market.* 17, 2/3 (2002), 107–118.
[28] Truman F. Bewley. 2007. *General Equilibrium, Overlapping Generations Models and Optimal Growth Theory.* Harvard University Press, Cambridge, MA, 8–16.
[29] Eric Benjamin Seufert. 2014. *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue.* Elsevier, Amsterdam, 22.
[30] Marius F. Niculescu and Dong Jun Wu. 2011. When should software firms commercialize new products via freemium business models. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.9580.
[31] Tim J. Smith. 2014. *Value-Based Pricing, Pricing Done Right: The Pricing Framework Proven Successful by the World's Most Profitable Companies.* Wiley-Blackwell, London, 11–34.
[32] Robert Harmon, David Raffo, Stuart Faulk. 2005. Value-based pricing for new software products: Strategy insights for developers. In *Proceedings of the Portland International Conference on Management of Engineering and Technology.*
[33] Anthony E. Boardman, David H. Greenberg, Aidan R. Vining, David L. Weimer. 2011. *Cost-Benefit Analysis: Concepts and Practice* (4th ed.). Prentice-Hall, Boston, MA, 27–32.
[34] Hugh M. Cannon and Fred W. Morgan. 1990. A strategic pricing framework. *J. Serv. Market.* 4, 2 (1990), 19–30.
[35] John L. Forbis and Nitin T. Mehta. 1981. Value-based strategies for industrial products. *Bus. Horiz.* 24, 3 (1981), 32–42.
[36] Andreas Hinterhuber. 2008. Customer value-based pricing strategies: Why companies resist. *J. Bus. Strategy* 29, 4 (2008), 41–50.
[37] Wedad Elmaghraby and Pinar Keskinocak. 2003. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Manage. Sci.* 49, 10 (2003), 1287–1309.
[38] Andrei M. Bandalouski, Mikhail Y. Kovalyov, Erwin Pesch, and S. Armagan Tarim. 2018. An overview of revenue management and dynamic pricing models in hotel business. *RAIRO-Operat. Res.* 52, 1 (2018), 119–141.
[39] Nayan B. Ruparelia. 2016. *Cloud Computing.* MIT Press, New York, NY, 17, 65.
[40] Benson Shapiro. 2002. Is Performance-based Pricing the Right Price for You? Retrieved from http://hbswk.hbs.edu/item/3021.html.
[41] Yu Hu, Jiwoong Shin, and Zhulei Tang. 2012. Performance-based pricing models in online advertising (unpublished).
[42] Retrieved from http://www.zdnet.com/article/amazon-web-services-marks-40th-price-drop-since-2006/.
[43] John Fernie, Suzanne Fernie, and Christopher Moore. 2015. *Principles of Retailing.* Routledge, 370.
[44] Rafael Becerril-Arreola, Mingming Leng, and Mahmut Parlar. 2013. Online retailers' promotional pricing, free-shipping threshold, and inventory decisions: A simulation-based analysis. *Eur. J. Operat. Res.* 230, 2 (2013), 272–283.
[45] Retrieved from https://aws.amazon.com/pinpoint/customer-engagement/customer-segmentation/.
[46] Asunción Mochón and Yago Sáez. 2015. *Understanding Auctions.* Springer, New York, NY, 25.
[47] Lawrence M. Ausubel and James J. Heckman. 2003. Auction theory for the new economy. In *New Economy Handbook.* Elsevier BV, North-Holland, 126–162.
[48] Philip Schofield. 2003. Jeremy Bentham: The principle of utility and legal positivism. *Current Legal Problems* 56, 1 (2003), 1.
[49] Rajkumar Buyya. 1999. *High-Performance Cluster Computing: Architectures and Systems, 1 and 2.* Prentice-Hall, Upper Saddle River, NJ, pp. 9–19.
[50] William H. Bell, David G. Cameron, Ruben Carvajal-Schiaffino, A. Paul Millar, Kurt Stockinger, and Floriano Zini. 2003. Evaluation of an economy-based file replication strategy for a data grid. In *Proceedings of the International Workshop on Agent-based Cluster and Grid Computing.* IEEE Computer Society Press.
[51] D. Rajkumar Buyya and Abramson S. Venugopal. 2005. The grid economy. *Proc. IEEE* 93, 3 (2005), 698–714.
[52] Rajkumar Buyya. 2009. Market-oriented cloud computing: Vision hype and reality of delivering computing as the 5th utility, In *Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid 2009 (CCGRID'09).*
[53] G. Kai. Hwang, C. Fox, and J. J. Dongarra. 2013. *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things.* Elsevier, 51.
[54] Rajkumar. Buyya, D. Abramson, J. Giddy, and H. Stockinger. 2002. Economic models for resource management and scheduling in grid computing. *Concurr. Comput.: Pract. Exper.* 14, 13–15 (2002), 1507–1542.

[55] Adel Nadjaran Toosi, Kurt Vanmechelen, Kotagiri Ramamohanarao, and Rajkumar Buyya. 2015. Revenue maximization with optimal capacity control in infrastructure as a service cloud markets, *IEEE Trans. Cloud Comput.* 3, 3 (2015), 261–274.
[56] Hong Xu and Baochun Li. 2013. Dynamic cloud pricing for revenue maximization. *IEEE Tran. Cloud Comput.* 1, 2 (2013), 158–171.
[57] Guillermo Gallego and Garrett van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Manage. Sci.* 40, 8 (1994), 999–1020.
[58] Retrieved from https://aws.amazon.com/blogs/aws/category/ec2-spot-instances/.
[59] Retrieved from https://moz.com/devblog/amazon-ec2-spot-request-volatility-hits-1000hour.
[60] Albert Greenberg, James Hamilton, David A. Maltz, and Parveen Patel. 2008. The cost of a cloud: Research problems in data center networks, *ACM SIGCOMM Comput. Commun. Rev.* 39, 1 (2008), 68–73.
[61] Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. 2013. Deconstructing amazon ec2 spot instance pricing. *ACM Trans. Econ. Comput.* 1, 3 (2013), 16.
[62] Liang Zheng, Carlee Joe-Wong, Chee Wei Tan, Mung Chiang, and Xinyu Wang. 2015. How to bid the cloud. *ACM SIGCOMM Comput. Commun. Rev.* 45, 4 (2015), 71–84.
[63] Artur Andrzejak, Derrick Kondo, and Sangho Yi. 2010. Decision model for cloud computing under SLA constraints. In *Proceedings of the 2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. 257–266.
[64] Michele Mazzucco and Marlon Dumas. 2011. Achieving performance and availability guarantees with spot instances. In *Proceedings of the 2011 IEEE 13th International Conference on High-Performance Computing and Communications (HPCC'11)*. 296–303.
[65] Qi Zhang, Quanyan Zhu, and Raouf Boutaba. 2011. Dynamic resource allocation for spot markets in cloud computing environments. In *Proceedings of the 4th IEEE International Conference on Utility and Cloud Computing (UCC'11)*. 178–185.
[66] Song Yang, Murtaza Zafer, and Kang-Won Lee. 2012. Optimal bidding in spot instance market. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'12)*. 190–198.
[67] ShaoJie Tang, Jing Yuan, and Xiang-Yang Li. 2012. Towards optimal bidding strategy for Amazon EC2 cloud spot instance. In *Proceedings of the 2012 IEEE 5th International Conference on Cloud Computing (CLOUD'12)*.
[68] Sangho Yi, Derrick Kondo, and Artur Andrzejak. 2010. Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud. In *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD'10)* 236-243.
[69] Junliang Chen, Chen Wang, Bing Bing Zhou, Lei Sun, Young Choon Lee, and Albert Y. Zomaya. 2011.Tradeoffs between profit and customer satisfaction for service provisioning in the cloud. In *Proceedings of the 20th International Symposium on High Performance Distributed Computing (HPDC '11)*. 229–238.
[70] Weichao Gao, Kang Chen, Yong wei Wu, and Weimin Zheng. 2015. Bidding for highly available services with low price in spot instance market. In *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing (HPDC'15)*. 191–202.
[71] Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, and Roger E. Bohn. 1988. *Spot Pricing of Electricity*. Springer, New York, NY, 32.
[72] Guofu Feng, Saurabh Garg, Rajkumar Buyya, and Wenzhong Li. 2012. Revenue maximization using adaptive resource provisioning in cloud computing environments. In *Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing*. 192–200 and 1550–5510.
[73] Retrieved from https://moz.com/blog/crawl-outage.
[74] Retrieved from https://blog.serverdensity.com/cloud-vs-colocation/.
[75] Caesar Wu and Rajkumar Buyya. 2015. *Cloud Data Centers and Cost Modeling: A Complete Guide to Planning, Designing and Building a Cloud Data Center*. Morgan Kaufmann, 167 and 690.
[76] Retrieved from https://www.energycouncil.com.au/analysis/worldwide-electricity-prices-how-does-australia-compare/.
[77] Edward Walker. 2009. The real cost of a CPU hour. *Computer* 42, 4 (2009), 35–41.
[78] Edward Walter, Brisken Walker, and Jonathan Romney. 2010. To lease or not to lease from storage clouds, *Computer* 43, 4 (2010), 44–50.
[79] Retrieved from https://hblok.net/blog/storage/.
[80] Hong Xu and Baochun Li. 2013. A study of pricing for cloud resources. *ACM SIGMETRICS Perf. Eval. Rev.* 40, 4 (2013), 3–12.
[81] Bernd Luderer, Volker Nollau, and Klaus Vetters. 2009. *Mathematical Formulas for Economists*. Springer Science & Business Media, Berlin,. 60–89.

[82] Mansaf Alam, Kashish Ara Shakil, Shuchi Sethi. 2015. Analysis and clustering of the workload in google cluster trace based on resource usage. In *Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE'15) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC'15) and 15th International Symposium on Distributed Computing and Applications for Business Engineering (DCABES'15)*. 740–747.
[83] Maho Nakata. 2012. All about RIKEN Integrated Cluster of Clusters (RICC), *Int. J. Netw. Comput.* 2, 2 (2012), 206–215.
[84] Siham El Kihal, Christian Schlereth, and Bernd Skiera. 2012. Price comparison for infrastructure-as-a-Service. In *ECIS*. 161.
[85] Persefoni Mitropoulou, Evangelia Filiopoulou, Christos Michalakelis, and Mara Nikolaidou. 2016. Pricing cloud IaaS services based on a hedonic price index, *Computing* 98, 11 (2016), 1075–1089.
[86] Liang Zhang. 2016. Price Trends for Cloud Computing Services. Retrieved from https://repository.wellesley.edu/thesiscollection/386/.
[87] Ariel Pakes. 2003. A reconsideration of hedonic price indexes with an application to PC's. *Am. Econ. Rev.* 93, 5 (2003), 1578–1596.
[88] Navendu Jain, Ishai Menache, Joseph (Seffi) Naor, and Jonathan Yaniv. 2014. A truthful mechanism for value-based scheduling in cloud computing. *Theory Comput.* 54, 3 (2014), 388–406.
[89] Draen Lucanin, Ilia Pietri, Ivona Brandic, and Rizos Sakellariou. 2015. A cloud controller for performance-based pricing. In *Proceedings of the 2015 IEEE 8th International Conference on Cloud Computing*. 155–162.
[90] Arpan Kumar Kar and Atanu Rakshit. 2014, Pricing of Cloud IaaS based on feature prioritization-A value-based approach. In *Recent Advances in Intelligent Informatics*. Springer, Cham, 321–330.
[91] Linlin Wu, Saurabh Kumar Garg, and Rajkumar Buyya. 2011. Sla-based resource allocation for software as a service provider (saas) in cloud computing environments. In *Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 195–204.
[92] Dave Durkee. 2010. Why cloud computing will never be free. *Queue* 8, 4 (2010), 20.
[93] Adam Eivy. 2017. Be wary of the economics of "serverless" cloud computing. *IEEE Cloud Comput.* 4, 2 (2017), 6–12.
[94] Retrieved from https://aws.amazon.com/lambda/pricing/.
[95] Peter Sbarski. 2017. *Serverless Architectures on AWS*, Manning Publications, New York, NY, 2–15.
[96] Retrieved from https://451research.com/blog/1351-application-containers-will-be-a-$2-7bn-market-by-2020,-representing-a-small-but-high-growth-segment-of-the-cloud-enabling-technologies-market.
[97] Robert J. Dolan. 2003. *Pricing: A Value-Based Approach*. Harvard Business School. 500–071.
[98] Stamatia Rizou and Ariana Polyviou. 2012. Towards value-based resource provisioning in the cloud. In *Proceedings of the 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. 155–160.
[99] Iwao Hirose and Jonas Olson. 2015. *The Oxford Handbook of Value Theory*. Oxford University Press, Oxford, 13.
[100] Retrieved from https://go.forrester.com/blogs/16-02-02-salesforce_announces_new_pricing_and_packaging_what_it_means_to_you/.
[101] N. Gregory Mankiw. 2018. *Principles of Economics*. Cengage Learning Australia, Melbourne, 28.
[102] I. A. Kash and P. B. Key. 2016. Pricing the cloud. *IEEE Internet Comput.* 20, 1 (2016), 36–43.
[103] NIST Cloud Computing Service Metrics Description. Retrieved from https://www.nist.gov/publications/cloud-computing-service-metrics-description.
[104] Oracle Service Cloud, Customer Experience Metrics and Key Performance Indicators. Retrieved from http://www.oracle.com/us/products/applications/cx-metrics-kpi-dictionary-1966465.pdf.
[105] Retrieved from https://storageservers.wordpress.com/2012/11/22/microsoft-proves-practically-that-vmware-is-too-expensive/.
[106] Retrieved from https://www.dedoimedo.com/computers/vmware-workstation-14.html.
[107] Caesar Wu, Rajkumar Buyya, and Ramamohanarao Kotagiri. 2018.Cloud computing market segmentation, In *Proceedings of the 13th International Conference on Software Technologies (ICSOFT 2018)*, ISBN: 978-989-758-320-9, Porto, Portugal.
[108] Joseph Keim Campbell, Michael O'Rourke, and Matthew H. Slater, eds. (Eds.). 2011. Carving nature at its joints: Natural kinds in metaphysics and science. MIT Press. http://www.jstor.org/stable/j.ctt5hhj54.
[109] Retrieved from https://iex.ec/.
[110] Retrieved from https://www.cloudandheat.com/.
[111] Retrieved from https://www.ringcentral.com/.
[112] Retrieved from https://databricks.com/.
[113] Caesar Wu, Adel Nadjaran Toosi, Rajkumar Buyya, and Ramamohanarao Kotagiri. 2018. Hedonic pricing of cloud computing services. *IEEE Trans. Cloud Comput.* 99, 1 (2018).
[114] Retrieved from https://cloud.google.com/tpu/docs/pricing.
[115] Carlee Joe-Wong and Soumya Sen. 2012. Mathematical frameworks for pricing in the cloud: Revenue, fairness, and resource allocations. https://arxiv.org/abs/1212.0022.

[116] Carlee Joe-Wong and Soumya Sen. 2013. Pricing the cloud: Resource allocations, fairness, and revenue. In *Proceedings of the Workshop on Information Technology & Systems (WITS'13)*.
[117] Chiang Mung, Steven H. Low, A. Robert Calderbank, and John C. Doyle. 2007. Layering as optimization decomposition: A mathematical theory of network architectures. *Proc. IEEE* 95, 1 (2007), 255–312.
[118] Soumya Sen, Carlee Joe-Wong, Sangtae Ha, and Mung Chiang. 2014. *Smart Data Pricing*. John Wiley & Sons, 127–166.
[119] Cinar Kilcioglu and Justin M. Rao. 2016. Competition on price and quality in cloud computing. In *Proceedings of the Conference of the World Wide Web (WWW'16)*. ACM.
[120] Mohammad Shahrad, Cristian Klein, Liang Zheng, Mung Chiang, Erik Elmroth, and David Wentzlaff. 2017. Incentivizing self-capping to increase cloud utilization. In *Proceedings of the 2017 Symposium on Cloud Computing*. 52–65.
[121] Mohammad Shahrad and David Wentzlaff. 2016. Availability knob: Flexible user-defined availability in the cloud. In *Proceedings of the 7th ACM Symposium on Cloud Computing*. 42–56.
[122] Roberts S. Fred. 1984. *Measurement Theory with Applications to Decision-Making Utility and the Social Sciences*. Cambridge University Press. 6–8.
[123] John L. Daly. 2002. *Pricing for Profitability: Activity-Based Pricing for Competitive Advantage*. John Wiley & Sons, Inc., 16.
[124] Shi Chen, Hau Lee, and Kamran Moinzadeh. 2019. Pricing schemes in cloud computing: Utilization-based vs. reservation-based. *Prod. Operat. Manag.* 28, 1 (2019), 82–102.
[125] V. K. Cody Bumgardner. 2016. *OpenStack in Action*. Manning Publications Company, 5.
[126] Retrieved from https://www.nist.gov/news-events/news/2011/10/final-version-nist-cloud-computing-definition-published.
[127] Venkat Ramaswamy and Kerimcan Ozcan. 2017. What is co-creation? An interactional creation framework and its implications for value creation. *J. Bus. Res.* 84 (2017), 196–205.
[128] Ian Morris. 2010. *Why the West Rules—For Now: The Patterns of History and What They Reveal about the Future*. Profile Books, 23.
[129] Paul Milgrom and Paul Robert Milgrom. 2004. *Putting Auction Theory to Work*. Cambridge University Press. 35.
[130] Paul Klemperer. 2004. *Auctions: Theory and Practice*. Princeton University Press, 11.
[131] Retrieved from https://aws.amazon.com/ec2/spot/.
[132] Sharrukh Zaman and Daniel Grosu. 2012. Combinatorial auction-based mechanisms for VM provisioning and allocation in clouds, In *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID'12)*. IEEE Computer Society, 729-734.
[133] David Porter, Stephen Rassenti, Anil Roopnarine, and Vernon Smith. 2003. Combinatorial auction design. *Proc. Natl. Acad. Sci. U.S.A.* 100, 19 (2003), 11153–11157.
[134] Peter C. Cramton, Yoav Shoham, and Richard Steinberg. 2006. *Combinatorial Auctions*. MIT Press, 30.
[135] Peter. W. Bridgman. 1959. The logic of modern physics' after thirty years. *Daedalus*, 88, 3 (1959), 518–526.
[136] Retrieved from https://aws.amazon.com/ec2/spot/instance-advisor/.

Received December 2018; revised June 2019; accepted June 2019