Journal of the Operational Research Society (2017) 68, 1030–1044

www.palgrave.com/journals

Optimal contract design for cloud computing service with resource service guarantee

Chia-Wei Kuo[1], Kwei-Long Huang[2] and Chao-Lung Yang[3]*

[1]*Department of Business Administration, National Taiwan University, 1 Sec.4, Roosevelt Road, Taipei 106, Taiwan;* [2] *Institute of Industrial Engineering, National Taiwan University, 1 Sec.4, Roosevelt Road, Taipei 106, Taiwan; and* [3] *Department of Industrial Management, National Taiwan University of Science and Technology, 43, Sec. 4, Keelung Road, Taipei 106, Taiwan*

The optimal contract design for cloud computing service with resource guarantee under the consideration of resource redundancy and network externality is studied in this research. A model in which a service provider determines joint pricing and resource allocation decisions is constructed by proposing two types of contracts with different service-level agreements (SLAs). The SLA of each contract describes the price and associated penalty if the provider cannot provide the resource requested by the customers. Optimal pricing and resource allocation decisions as well as the equilibrium contracts of the service provider are analyzed based on the dynamics of the model characteristics. We found that optimal contract design is sensitive to both service levels and customers' beliefs of compensation ratio when the requested resource is unfulfilled. Furthermore, service providers should evaluate the trade-off between benefit of price discrimination and effect of network externality when determining the optimal contract design.

Journal of the Operational Research Society (2017) **68(9),** 1030–1044. doi:10.1057/s41274-016-0141-z; published online 21 December 2016

Keywords: cloud computing service; pricing; SLA; contract design; service guarantee

1. Introduction

Cloud computing is a system that offers computing resources such as computational capacity, storage, and applications as services by information technology infrastructures over the Internet to minimize management effort. The evolution of new forms of data communication and the development of a highly capable Internet infrastructure allowed cloud computing providers such as Amazon, Google, eBay, and Microsoft to deliver a variety of services such as infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) over the Internet to their customers, or provide system hardware and software in data centers not only for corporate operations but also for personal use. In 2015, International Data Corporation (IDC) has announced a report that forecasts that the worldwide public IT cloud services revenue will increase from $96.5 billion in 2016 to $195 billion in 2020.[1] IDC's white paper in November 2012 also reports the demand of "cloud-savvy" IT workers will grow by 26% annually through 2015[2]. In addition, a report in the *New York Times* in 2012 indicates that small start-up companies and large corporations utilize computing services supported by cloud computing such as the Amazon Web Service (AWS) infrastructure. Several Internet service providers such as Netflix utilize cloud computing to provide fast and high-quality service to its customers. Thomson Reuters also reports that the company's webcasting operating expenses declined by 40–50% owing to the use of the cloud computing platform.

In order to maintain ubiquitous accessibility of data, cloud computing service providers need to ensure reliable resource management and provide high standard of the storage, network, and hardware availability to their customers. The redundancy of hardware is essential to reduce the risk of service discontinuity caused by the hardware or network failure. The information technology of the hardware hot swapping or plugging allows the redundant hardware or network device to attach in the service line in a short time once the failure occurs. In addition to the resource redundancy, cloud computing service providers also adjust their resource to handle the resource request by the network externality effect. Positive network externality exists if the benefit of offering

*Correspondence: Chao-Lung Yang, Department of Industrial Management, National Taiwan University of Science and Technology, 43, Sec. 4, Keelung Road, Taipei 106, Taiwan.
E-mail: clyang@mail.ntust.edu.tw

[1]Source: https://www.idc.com/getdoc.jsp?containerId=prUS41669516.

[2]Source: http://cloudtimes.org/2013/01/09/idc-report-1-7-million-cloud-computing-jobs-remain-unfilled-gap-widening/.

services to customers is an increasing function of the number of other users. For example, the online data synchronization cloud computing service such as Dropbox can gain more adhesive customer usage when more users join the network. On the other hand, negative network externality exists if the benefits are a decreasing function of the number of other users. When more users join the network, the cloud computing service provider needs to invest more hardware to maintain the service. If the service is provided by the unlimited basis rather than the usage basis, such as the unlimited 4G data plan in cell phone service, the network externality might cause negative utility and the system burden tends to increase the possibility of unstable service. Therefore, from the point of view of the resource management, the trade-off between offering the stable service and enlarging the usage body is an interesting characteristic of service maintenance in cloud computing.

When offering cloud service to the end customers, cloud service providers need to evaluate the existing resource and the associated cost structures so as to design the service contracts. Obviously, a contract between a cloud computing service provider and a customer specifies the desired services of the customer. The service provider attempts to simultaneously fulfill the needs of the customer and achieve maximum profit. A service-level agreement (SLA) is usually defined in the service contract as a guarantee of the number of available resource units such as speed, bandwidth, and storage space. If a service guarantee is not fulfilled, then the service provider would compensate by offering customers a price discount. For example, Amazon Elastic Compute Cloud (EC2) offers an SLA stating that the annual uptime percentage, the percentage of service availability per year, of Amazon EC2 should be at least 99.95%; otherwise, the customer will receive service credit for their payment. Similarly, Amazon Simple Storage Service (S3) also provides an SLA with monthly uptime percentage on storage availability of at least 99.9% during any monthly billing cycle; otherwise, 10% to 25% service credit will be provided to customers based on the availability percentage. A service contract with different SLAs represents different quality levels of the services perceived by the customers, and customers are willing to pay a higher price for a service contract with higher-quality level and better SLA if the associated price for the service is acceptable. Faced with different resource and service needs, researchers and practitioners are seeking paradigms to design service contracts that better fulfill the resource needs of end market so as to maximize the profit.

In recent years, market segmentation and price differentiation are gradually utilized in retailing (Bitran and Mondschein, 1997); however, only a few studies in literature have adopted price differentiation in the cloud computing environment, particularly in investigating joint pricing and resource allocation decisions under redundancy and network externality effect. In fact, a cloud service provider is able to differentiate markets by designing a combination of service categories with different SLAs and associated prices. For providers who offer cloud computing service, the question is how to establish prices and SLAs with a variety of penalties (price discount) in the service contract. Another question that may arise is how customers respond and select a preferred contract based on resource needs when faced with different service contracts. Our research intends to answer these questions by investigating how the service provider, upon knowing the best response of each customer, can reactively design the best service contract portfolio (i.e., the types of service contracts to be served) to maximize profit under various resource constraints.

In this research, we consider a model in which a cloud computing service provider with a limited units of resource can offer two types of service contracts, namely premium and basic contracts, with different prices and associated SLAs. Each SLA describes the various penalties that the service provider needs to compensate for if the resource requested by the customers cannot be fulfilled. Due to the nature of cloud computing, the service availability needs to consider the factor of the resource adjustment caused by the hardware redundancy and usage of network externality. The service provider needs to not only maximize the profit but also manage their limited resource to provide the acceptable SLA when facing possible system failure and influence by network externality. In the proposed model, customers differ not only in their willingness to pay but also in the number of resource units they request. By noting the prices and SLAs indicated in the contracts, customers form a belief of compensation for each contract, representing the expected compensation ratio to be paid by the service provider if the requested resource is not fulfilled. Customers then select a contract that best fits their individual resource needs if the price charged is below what they are willing to pay. After the contract is signed, the unit of resource requested by each customer is provided. If the requested units of resource are not provided owing to the limited resource of the provider and the rule of resource allocation, a price discount will be offered to the customer based on the SLA. As a result, the service provider decides to his best interest whether to provide a *dual contract* (both premium and basic contracts) or only a single contract (*single premium* or *single basic* contract) so as to maximize expected profit.

In this work, the service provider's profit functions are constructed in the proposed model for dual and single premium (or single basic) contract cases. Optimal prices are derived for each case depending on the unit of resource acquired by the service provider. The resource allocation rule is also discussed. A case in which the unit of resource is sufficiently large is considered to obtain further managerial insights. In this case, the conditions where a dual contract is offered rather than a single premium or basic contract are also determined. The results show that when the two service contracts are highly differentiated and/or the beliefs of compensation ratio for two contracts are moderate and close to each other, the service provider is better offering a dual contract. The prices of both contracts can be raised to increase the provider's profit. Optimal decision and

associated contract price are sensitive to the differences between the two service contracts and between the beliefs of compensation for the two contracts even though only a single contract is offered (i.e., single premium or single basic contract).

Our numerical study also provides some interesting findings. First, the profit of the service provider increases when the quality of each service is enhanced. Second, forming a high belief of compensation from the customers' point of view helps the provider establish better market price discrimination. However, such power is mitigated when the agreed penalty ratio described in the SLAs is high. We also find that a single premium contract is offered when the service provider possesses large units of resource; the opposite is true for a single basic contract. When the unit of resource is modest, offering a dual contract is appropriate for the service provider to balance the profits from two different market segments. In addition, although service provider benefits from price discrimination by offering a dual contract, this benefit weakens as the effect of negative network externality increases. This trade-off between market segmentation and network externality is especially significant for the decision makers in the cloud computing service environment.

The remainder of this paper is organized as follows. Section 2 provides a survey of relevant literature. Section 3 provides a description of our model and the rule of resource allocation. The analytical results and numerical study are presented in Sections 4 and 5. Sections 6 and 7 consider the effect of network externality and several extensions. The summary and managerial implications of this research are provided in Section 8.

2. Literature review

Revenue management in cloud computing services has received increasing attention owing to the advancement of mobile infrastructures and increased prevalence of various applications. Guan *et al* (2008) utilized an auction model for service providers to determine service levels and associated prices under the constraint of quality of service (QoS). Fulp and Reeves (2004) considered bandwidth provision and connection management with the objective of maximizing profit given the user demand estimation and connection duration. Zhang *et al* (2008) and Zhang *et al* (2009) discussed how two data service providers maximize their profit under duopoly pricing with delay guarantee. Jain and Kannan (2002), Sundararajan (2004), and Candogan *et al* (2012) considered the pricing problem of digital goods. Hosanagar *et al* (2005) analyzed pricing and capacity allocation policies for best-effort and premium-caching services. Bhargava and Sun (2008) studied performance-contingent pricing for Internet access services. Liu *et al* (2010) studied pricing policies for in-demand IT services with multiple service levels. Ganesh *et al* (2007) utilized game theory for congestion pricing by creating a communication network model with bandwidth sharing.

Resource allocation is also an important issue in cloud computing. Anandasivam and Premm (2009) utilized auction model and dynamic pricing to decide whether to accept a customer's request and the associated price with sharing one type of resources. Mihailescu and Teo (2010) employed dynamic pricing to increase user utility and acceptance of requests where resources offered by several providers. The integration can increase the scalability and reliability of clouds. Teng and Magoules (2010) applied game theory to address pricing and resource allocation in which customers "have" budgets and service deadlines. Thomas *et al* (2002) used admission control to allocate network resources for services of multiple types. Lin *et al* (2010) proposed a second-price auction mechanism for capacity allocation in cloud computing. An *et al* (2010) investigated a dynamic resource allocation problem where service providers and customers negotiate price. Beloglazov *et al* (2012) explored resource allocation in cloud computing for data centers with considering electrical energy consumption. Aloi *et al* (2012) adopted the inventory model to form rules of bandwidth allocation for wireless communication services. Our study integrates pricing and resource allocation for multiple cloud computing services to allow service providers to maximize profit while ensuring service quality.

Quality of service (QoS) is one of main characteristics of cloud computing that ensures the reliability for customers. Several researches consider QoS in their models including Zhang *et al* (2008, 2009); Hosanagar *et al* (2005); and Bhargava and Sun (2008). Rouskas *et al* (2008), Guan *et al* (2008), and Fulp and Reeves (2004) considered profit-maximizing problem under different QoS. Jukic *et al* (2004) considered two types of network services: network service with bandwidth and delay guarantee and best-effort network service without quality guarantee. Wei *et al* (2010) adopted a game-theoretic method to schedule cloud computing services with QoS. Fishburn and Odlyzko (2000) compared three network configurations with providing differential QoS. Ou *et al* (2006) utilized a queuing-based model to analyze a two-level web QoS system for bandwidth allocation and traffic congestion management.

Service-level agreements (SLAs) ensure quality of service and improve service satisfaction for users. Offering SLAs can also enhance the service quality and performance of a firm and provide the firm a competitive edge (Bhargava and Sun, 2008). The issues in SLA provision have been extensively studied in the areas of telecommunication, networking, and wireless networking. Fawaz *et al* applied it to optical domain (O-SLA) (Fawaz *et al*, 2004), and SLA is also used in evaluating cloud computing (Wu and Buyya, 2010). Ardagna *et al* (2007) investigated an SLA optimization problem in which a centralized network controller has to allocate requested applications to different servers and schedule workload for each server. In this present study, two types of service levels are also considered; an SLA is specified for each service. Given a single network or pool of computing resource, bandwidth and resource allocation need to be addressed when

considering multiple service levels. Kasap *et al* (2007) investigated a network resource acquisition problem where a firm minimizes the cost of acquiring capacity from several providers with different levels of service quality.

Network externality is the effect that the utility of a service for a user is affected by the number of other users (Shapiro and Varian, 1998). Kate and Shaprio (1994) indicated several strategies to attract users to networks such as low price, future commitment, and service reputation. For a cloud computing service, low price (even free charge) and the commitment of service sustainability are two common methods to attract users. Due to the high cost in the early stage of forming service network, the cloud computing service provider faces the challenge of utilizing resources with high SLA service and achieving profitability of running the business. Keskin and Taskin (2014) proposed a pricing model of cloud computing service focusing on expanding the user set with time-inconsistent behavior. Their works showed that the effect of network externality reduces the impact of low switching costs and the monopolist benefits from time-inconsistent behavior.

3. Model

In the proposed model, a cloud service market in which a service provider with T units of resource offers two types of service contracts, namely premium and basic service contracts, with different SLAs to end customers is considered. Each agreement represents a certain degree of guarantee on the fulfillment of the requested resource and the associated penalty to be paid by the service provider to the customers if such service guarantee cannot be fulfilled. Compared with basic contracts, premium contracts provide customers a high priority of acquiring the resource; that is, if the resource is limited and is not enough to satisfy all customer requirements, the customers with premium contracts are prioritized over those with basic contracts. Therefore, premium service contract is normally regarded as the contract with high quality. We denote the quality of premium and basic service contracts as q_H and q_L, respectively, where $0 < q_L < q_H \leq 1$. Premium and basic service contracts are sold by the service provider at the price of P_H and P_L, respectively. If the requirements of the customers are not fulfilled, the service provider will pay a penalty cost for violating the agreement in the contract. This penalty cost may differ depending on the type of the contract and level of unfulfilled requests. Besides penalty costs, the service provider also considers resource redundancy for premium service contract customers to secure the availability of the data access. Let $\delta \in [0, 1]$ be the redundancy factor of cloud resource. If a premium contract customer requires one unit of resource, the service provider needs to reserve $(1 + \delta)$ units of resource, among which δ units are for redundancy consideration. This resource redundancy consideration, however, does not apply to basic contract. Note that we have this setting to reflect the fact that premium contract represents a higher service to the customers and the service provider should keep more resource. One may consider a scenario under which the same redundancy is also applied to basic contract. Including this in our model influences the optimal decisions of the service provider; however, all managerial insights remain valid.

The sequence of events is as follows. The service provider first offers two types of service contracts and associated prices. Each customer selects the preferred contract service that provides him the highest gross utility. When selecting the contract, customers are unaware of exactly how many units of resource they will consume, which is a random variable from the viewpoint of both parties. A customer who accepts either a premium or basic contract is assumed to consume 1 or 2 units of resource with probability g_{i1} and g_{i2} to simplify the exposition where $i = H$ represents premium contract and $i = L$ represents basic contract. We have $\sum_{j=1}^{2} g_{Hj} = 1$ and $\sum_{j=1}^{2} g_{Lj} = 1$. Probabilities, g_{Hj} and g_{Lj}, are information known to the service provider based on historical data or prediction by experts. However, each individual customer cannot access such information. This private information helps the service provider determine the types of contracts to serve so as to better price discriminate the end market.

When customers use cloud service, the service provider may not be able to fulfill the units of resource requested by each customer, especially for customers who select the basic service contract owing to low priorities. Consider the customers who select the premium (or basic) contract and who pay price P_H (or P_L). If a customer requests one unit of resource and the service provider is unable to fulfill such request, the customer will receive a compensation of $\alpha_1 P_H$ (or $\alpha_3 P_L$) where $\alpha_1, \alpha_3 \in [0, 1]$. If a customer requests two units of resource and only one unit of resource is provided, a compensation of $\alpha_2 P_H$ (or $\alpha_4 P_L$) will be paid to the customer where $\alpha_2, \alpha_4 \in [0, 1]$. If none of the requested resources is allocated, the service provider has to pay $(\alpha_1 + \alpha_2)P_H$ (or $(\alpha_3 + \alpha_4)P_L$) to the customer. We assume that $\alpha_4 \leq \alpha_3 \leq \alpha_2 \leq \alpha_1$ to represent the fact that premium contract customers are more valuable to the service provider due to high margin. Also the provider is penalized more for not being able to provide the first unit of resource under either contract. One can interpret that the first unit of resource provides connectivity to the cloud service, and the second unit of resource determines the bandwidth or speed of cloud service connectivity. From the customer's point of view, the ability to access cloud service is more essential than bandwidth and we normalize the resource need of connectivity or bandwidth to a unit of resource. Here, we assume the service provider is penalized more for not being able to satisfy the second unit of resource under premium contract compared to the first unit of resource under basic contract. In Section 7, we consider one extension by assuming $\alpha_2 \leq \alpha_3$, namely satisfying the first unit of resource under premium or basic contract is more important than the second unit of resource.

Notice that when signing a contract with the service provider, the customer does not possess any information

regarding the probability g_{ij} where $i \in \{H, L\}, j = 1, 2$. A customer is thus unable to correctly identify whether his or her requests can be fulfilled and if not, what amount of compensation he or she will obtain. A customer simply forms a belief of obtaining the compensation from the service provider when agreeing to the contract. Let M_H and M_L be, respectively, the belief of the compensation ratios for the customers who select premium and basic service contracts for $0 \le M_L, M_H \le 1$. A customer chooses premium (or basic) contract pays P_H (or P_L) and expects to receive a compensation of $M_H P_H$ (or $M_L P_L$) if the requested resource cannot be fulfilled. The beliefs may be based on previous transactions with the service provider or on the forecast of their future resource requirement.

The customers are differentiated based on their willingness to pay. From the perspective of the service provider, the customer's willingness to pay, v, is a random variable. We assume that v follows a uniform distribution over interval [0, 1], as commonly assumed in economics, information systems, and marketing literature. This one-dimensional variable allows us to rank different services; each customer benefits from high gross utility when high quality is offered along this "more-is-better" quality dimension. In determining whether the customers select the premium or basic service contract, vq_i is utilized to denote the gross utility that the customer obtains from the service contract where q_i refers to the service quality for $i \in \{L, H\}$. A customer with willingness to pay v will select a service contract only if $V_i = vq_i - P_i + M_i P_i > 0$ where i stands for L and H in the service contract and $M_i P_i$ represents the expected compensation the customer will obtain if the service agreement cannot be fulfilled. Furthermore, a customer will select the premium service contract if $\max\{V_L, V_H, 0\} = V_H$. A customer will select the basic service contract if $\max\{V_L, V_H, 0\} = V_L$. If selecting either contract provides negative utility, the customer will select neither.

Based on the aforementioned information, customers can be distinguished into three market segments: selecting the premium service contract, selecting the basic contract, and signing no contract. The thresholds for market segmentation are derived by considering the customers' gross utility. We first consider the group of customers who selects the basic service contract. The customers in this segment believe that the basic service is beneficial and yields positive utility; otherwise, customers have no intention of availing the service. Thus, the willingness to pay threshold can be derived when the gross utilities of selecting the basic service and not buying are indifferent. We let the threshold between basic and non-buying be $\underline{v}$, and we obtain

$$\underline{v} q_L - P_L + M_L P_L = 0, \quad \text{or} \quad \underline{v} = \frac{(1 - M_L) P_L}{q_L}. \tag{1}$$

With regard to the selection of basic or premium service contracts, customers simply compare the gross utility of both contracts and select the one that provides higher gross utility. Therefore, the willingness to pay threshold (denoted as $\bar{v}$) is obtained by equating utilities of selecting the basic (with quality level q_L) and premium (with quality level q_H) service contracts. Hence, $\bar{v}$ can be expressed as

$$\bar{v} q_L - P_L + M_L P_L = \bar{v} q_H - P_H + M_H P_H, \quad \text{or}$$
$$\bar{v} = \frac{(1 - M_H) P_H - (1 - M_L) P_L}{q_H - q_L}. \tag{2}$$

Based on the thresholds of willingness to pay, $\bar{v}$ and $\underline{v}$, the market can be divided into three segments: above $\bar{v}$, between $\bar{v}$ and $\underline{v}$, and below $\underline{v}$. By comparing willingness to pay with the thresholds, customers are identified as those who select the premium service, the basic service, and no service. The market share for basic and premium service contracts is $\bar{v} - \underline{v}$ and $1 - \bar{v}$, respectively. Recall that g_{Hj} and g_{Lj} are the probabilities of requiring j units of resource from both contracts. Let X_{Hj} and X_{Lj} be the respective market share of requiring j units of resource when selecting premium and basic service contracts for $j = 1, 2$. We obtain $X_{Hj} = (1 - \bar{v}) g_{Hj}, X_{Lj} = (\bar{v} - \underline{v}) g_{Lj}$, where $j = 1, 2$.

In addition to offering premium and basic service contracts simultaneously, the provider can also opt to provide only one contract: either single premium or single basic contract. The service provider simplifies the service content and focuses on resource fulfillment. In the case where the service provider offers only a single premium (or basic) contract, we let v_p (or v_b) be the willingness to pay threshold above which the customers select the premium (or basic) contract and below which the customers do not purchase when only a premium (or basic) contract is provided at the price of P_H (or P_L). We then obtain $v_p = \frac{(1 - M_H) P_H}{q_H}$ (or $v_b = \frac{(1 - M_L) P_L}{q_L}$). Therefore, market shares are, respectively,

$$X_{H1} = \left(1 - \frac{(1 - M_H) P_H}{q_H}\right) g_{H1}, \quad X_{H2} = \left(1 - \frac{(1 - M_H) P_H}{q_H}\right) g_{H2}$$
$$X_{L1} = \left(1 - \frac{(1 - M_L) P_L}{q_L}\right) g_{L1}, \quad X_{L2} = \left(1 - \frac{(1 - M_L) P_L}{q_L}\right) g_{L2}. \tag{3}$$

3.1. Allocation rule

The service provider offers a dual contract (both premium and basic contracts), and penalty costs will be paid to the customers if the units of resource requested by the customers cannot be fulfilled. Given the assumption that $\alpha_4 \le \alpha_3 \le \alpha_2 \le \alpha_1$, fulfilling the first unit of resource to customers who sign the premium contract has the highest priority followed by offering the second unit of resource to customers who sign the premium contract and request two units of resource. The remaining resource will be assigned to customers who select the basic contract for the same logic. Hence, the allocation rule adopted by the service provider follows the order: (i) Offer one unit of resource to each customer who signs the premium contract whether the

customer requests one or two units of resource;[3] (ii) Offer one unit of resource to customers who sign the premium contract and request two units of resource to fulfill the second unit of resource required; (iii) Offer one unit of resource to customers who sign the basic contract whether the customers request one or two units of resource; and (iv) Offer one unit of resource to customers who sign the basic contract and request two units of resource.

When single premium contract (or single basic contract) is offered, the allocation rule basically adheres to (i) and (ii) (or (iii) and (iv)) above. Based on the allocation rule mentioned above, the problems brought about by the service provider offering both premium and basic contracts can be separated into five different cases depending on resource, T. The proportion of the customers who sign the premium contract ($i = H$) or the basic contract ($i = L$) and request j units of resource is denoted by X_{ij}, where $i \in \{H, L\}, j = 1, 2$. The case wherein premium and basic contracts are offered simultaneously is first considered.

- **Case 1**: When T is sufficiently small: $T \le (1+\delta)(X_{H1} + X_{H2})$
  In this case, the service provider can only accommodate a certain number of premium contract customers given the requirement of resource redundancy. Based on the allocation rule, the service provider will offer one unit of resource to premium contract customers whether the customers request one or two units of resource. Since $T \le (1+\delta)(X_{H1} + X_{H2})$, $\frac{(1+\delta)(X_{H1}+X_{H2})-T}{1+\delta}$ premium contract customers cannot obtain the first unit and all premium contract customers who request two units are not fulfilled. Furthermore, none of the basic contract customers are able to obtain the unit of resource. Therefore, $\left(\frac{(1+\delta)(X_{H1}+X_{H2})-T}{1+\delta}\alpha_1 + X_{H2}\alpha_2\right)P_H$ are paid to premium customers and $(X_{L1}\alpha_3 + X_{L2}(\alpha_3+\alpha_4))P_L$ are paid to basic contract customers. The service provider's profit, π_B^1, can be obtained as follows:

$$\begin{aligned}\pi_B^1 =& (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L \\ &- \left(\frac{(1+\delta)(X_{H1}+X_{H2})-T}{1+\delta}\alpha_1 + X_{H2}\alpha_2\right)P_H \\ &- (X_{L1}\alpha_3 + X_{L2}(\alpha_3+\alpha_4))P_L,\end{aligned} \tag{4}$$

  where the first two terms represent the revenue from both contract customers and the last two terms are the penalties paid to premium and basic contract customers, respectively. Hence, the service provider's problem is selecting prices, P_H and P_L, to maximize the following problem: $\max_{\{0<P_L\le P_H\}} \pi_B^1$ subject to $T \le (1+\delta)(X_{H1} + X_{H2})$.

- **Case 2**: When T is relatively small: $(1+\delta)(X_{H1} + X_{H2}) \le T < (1+\delta)(X_{H1} + 2X_{H2})$
  When $(1+\delta)(X_{H1} + X_{H2}) \le T$, all premium contract customers can obtain their first requested unit of resource regardless of whether how many units they request. The rest of the resource will be utilized to supply the second unit of resource to avoid a penalty of $\alpha_2 P_H$. Therefore, the service provider chooses P_H and P_L to maximize the profit π_B^2 subject to $(1+\delta)(X_{H1} + X_{H2}) \le T < (1+\delta)(X_{H1} + 2X_{H2})$, where

$$\begin{aligned}\pi_B^2 =& (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L \\ &- \left(\frac{(1+\delta)(X_{H1}+2X_{H2})-T}{1+\delta}\right)\alpha_2 P_H \\ &- (X_{L1}\alpha_3 + X_{L2}(\alpha_3+\alpha_4))P_L.\end{aligned} \tag{5}$$

- **Case 3**: When T is modest: $(1+\delta)(X_{H1} + 2X_{H2}) \le T < (1+\delta)(X_{H1} + 2X_{H2}) + X_{L1} + X_{L2}$
  Considering that $(1+\delta)(X_{H1} + 2X_{H2}) \le T$, the service provider is able to fulfill all the units requested by premium contract customers. Furthermore, $T < (1+\delta)(X_{H1} + 2X_{H2}) + X_{L1} + X_{L2}$ represents the fact that only some of the units requested by basic contract customers are satisfied regardless of whether the customers request one or two units of resource. Therefore, the service provider sets prices P_H and P_L to maximize the profit

$$\begin{aligned}\pi_B^3 =& (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L \\ &- (((1+\delta)(X_{H1}+2X_{H2}) \\ &+ X_{L1} + X_{L2} - T)\alpha_3 + X_{L2}\alpha_4)P_L\end{aligned}$$

  with the constraint

$$\begin{aligned}&(1+\delta)(X_{H1} + 2X_{H2}) \le T < (1+\delta)(X_{H1} + 2X_{H2}) \\ &+ X_{L1} + X_{L2}.\end{aligned} \tag{6}$$

- **Case 4**: When T is relatively large: $(1+\delta)(X_{H1} + 2X_{H2}) + X_{L1} + X_{L2} \le T < (1+\delta)(X_{H1} + 2X_{H2}) + X_{L1} + 2X_{L2}$
  Similarly, in this case, basic contract customers only obtain one unit of resource and some of basic contract customers who request two units will receive compensation. The service provider's profit is

$$\begin{aligned}\pi_B^4 =& (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L \\ &- ((X_{H1}+2X_{H2})(1+\delta) + X_{L1} + 2X_{L2} - T)\alpha_4 P_L.\end{aligned} \tag{7}$$

- **Case 5**: When T is sufficiently large: $(1+\delta)(X_{H1} + 2X_{H2}) + X_{L1} + 2X_{L2} \le T$
  If the resource of the service provider is sufficiently large, the service provider simply provides the requested units of customers who sign both premium and basic contracts. Here, the service provider's profit is

$$\pi_B^5 = (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L, \tag{8}$$

  and no penalty is incurred.

[3] If the resource cannot be allocated to each customer who requests one unit of resource, the resource is randomly allocated to the customers and the customer will obtain the unit of resource with equal probability. The same rule when the unit of resource is inadequate applied to (ii), (iii), and (iv) as well.

Given the unit of resource, T, the service provider solves the five cases separately to determine optimal prices P_H and P_L in each case. The service provider then compares the profit that can be obtained from the cases to determine the optimal prices to be adopted.

The case where the service provider offers either single premium or single basic contract is also analyzed. Each contract provides three different cases depending on the relationship between resource T and proportion of customers who sign the contract, X_{ij} where $i \in \{L, H\}$ and $j = 1, 2$. The provider determines either P_H or P_L to maximize the profit given that the price satisfies the associated constraint in each case. The outcomes are summarized in Table 1.

4. Analysis

The optimal prices set by the service provider depending on whether a dual contract or only one contract (i.e., single premium or single basic) is offered to the customers are discussed in this section. We analyze the service provider's optimal pricing decisions depending on whether the unit of resource is sufficiently large or not. In this section we assume $g_{H1} = g_{L1} = g_1$ and $g_{H2} = g_{L2} = g_2$ to represent the fact that the proportion of customers requesting one or two units of resource does not depend on the contract selected by the customers.

4.1. When unit of resource is sufficiently large

In this subsection, we consider a circumstance where the service provider acquires a sufficiently large resource T to obtain managerial insights. This coincides with practical instances such as cloud computing hadoop platform and a scalable parallel computing infrastructure where the computational node can be easily expanded and upgraded in a flexible manner with commodity machines. Therefore, we focus on the case of $(1+\delta)(X_{H1}+2X_{H2}) + X_{L1} + 2X_{L2} \le T$ when both contracts are offered and on the case of $(1+\delta)(X_{H1}+2X_{H2}) < T$ or $X_{L1} + 2X_{L2} < T$ when single premium or single basic contract is offered. The following proposition summarizes the optimal price(s) for each contract.

Proposition 1 *If the service provider offers the dual contract (both premium and basic contracts), the optimal prices are* $P_H^d = \frac{2N_L q_H (q_H - q_L)}{4N_H N_L q_H - (N_H + N_L)^2 q_L}$ *and* $P_L^d = \frac{q_L (N_H + N_L)(q_H - q_L)}{4N_H N_L q_H - (N_H + N_L)^2 q_L}$. *If the service provider offers single premium or single basic contract, the respective optimal prices,* $P_H^s = \frac{q_H}{2(1-M_H)}$ and $P_L^s = \frac{q_L}{2(1-M_L)}$ *where* $N_H = 1 - M_H$ *and* $N_L = 1 - M_L$.

Given the optimal prices for each contract offered by the service provider, we then discuss the effects of model characteristics on the optimal price(s) of each contract. The following corollaries posit the conditions where the service provider offers both (dual contract) or only one of the contracts (single premium or single basic contract).

Corollary 1 *Consider the service provider offers the dual contract (both premium and basic contracts) and* P_H^d *and* P_L^d *are the prices of premium and basic contracts, respectively:*

(a) P_H^d *increases in* M_H, P_H^d *decreases in* M_L *if* $M_L < M_H$,
(b) P_L^d *increases in both* M_H *and* M_L *if* $\frac{(M_H + M_L - 2)^2}{4(1-M_L)^2} < \frac{q_H}{q_L}$,
(c) P_H^d *increases in* q_H *if* $\frac{(M_H + M_L - 2)^2}{4(1-M_H)(1-M_L)} < \frac{q_H^2}{q_L(2q_H - q_L)}$, *and* P_L^d *decreases in* q_H, *and*
(d) *both* P_H^d *and* P_L^d *increase in* q_L.

When the service provider offers both contracts simultaneously, the price for the premium contract, P_H^d, is found to increase in the belief of compensation M_H. Also, P_H^d decreases in the belief of compensation M_L if $M_L < M_H$. However, the same conclusions may not apply to the price for the basic contract, P_L^d. When both services are fully differentiated (i.e., larger q_H/q_L), the service provider is allowed to segment the market aggressively without worrying about the cannibalization effect wherein customers expect to receive large compensation from a premium contract (i.e., high M_H). The price for the basic contract can be increased in the aim of acquiring a high profit margin. However, a high M_L would mitigate this price differentiation because high compensation belief for the basic contract causes the selection of low-end service, which indirectly reduces the difference between the two prices. When the quality levels of the two services are close (i.e., smaller

Table 1 Profit of service provider for single premium and single basic contracts

Contract	*Case (constraint)*	*Provider's profit*
Premium	$T < (1+\delta)(X_{H1} + X_{H2})$	$(X_{H1} + X_{H2})P_H - \left(\frac{(1+\delta)(X_{H1}+X_{H2}) - T}{1+\delta}\alpha_1 + X_{H2}\alpha_2\right)P_H$
	$(1+\delta)(X_{H1} + X_{H2}) \le T < (1+\delta)(X_{H1} + 2X_{H2})$	$(X_{H1} + X_{H2})P_H - \frac{(1+\delta)(X_{H1}+2X_{H2}-T)}{1+\delta}\alpha_2 P_H$
	$(1+\delta)(X_{H1} + 2X_{H2}) \le T$	$(X_{H1} + X_{H2})P_H$
Basic	$T < X_{L1} + X_{L2}$	$(X_{L1} + X_{L2})P_L - ((X_{L1} + X_{L2} - T)\alpha_3 + X_{L2}\alpha_4)P_L$
	$X_{L1} + X_{L2} \le T < X_{L1} + 2X_{L2}$	$(X_{L1} + X_{L2})P_L - (X_{L1} + 2X_{L2} - T)\alpha_4 P_L$
	$X_{L1} + 2X_{L2} \le T$	$(X_{L1} + X_{L2})P_L$

q_H/q_L), the prices of both contracts follow different patterns with respect to M_H and cause a decrease in M_L.

Parts (c) and (d) of the corollary show that the service provider tends to increase the prices of both services when the quality level of the basic service (q_L) increases. Such result is expected because the service provider can charge a high price for the basic contract owing to the enhancement of the service and increase the premium contract price accordingly to easily distinguish both contracts. An increase in the quality of the premium contract does not necessarily induce high premium contract price. When the two services are not sufficiently differentiated, the service provider needs to reduce both prices to balance the profits by two services.

Corollary 2 *Consider the service provider offers single premium (or basic) contract with P_H^s (or P_L^s), then P_H^s increases in both q_H and M_H and P_L^s increases in both q_L and M_L.*

In addition, the effects of M_H, M_L, q_H, and q_L on optimal price when a single premium or basic contract is offered are also analyzed. Optimal price should increase with the belief of compensation because customers are willing to pay a high price when they believe that a large portion of the compensation will be returned to them if the service provider fails to fulfill their needs. Customers expect a high price when the quality level of the service is enhanced. Therefore, both P_H^s and P_L^s increase in their respective quality level, q_H and q_L, accordingly. In the case when the service provider offers only single premium or basic contract, she does not need to take into account the negative effect from the other contract upon setting the optimal price and thus, the scenario is simpler compared to the dual contract.

The profits of the service provider under different service contracts are compared. The following proposition investigates the trade-off between offering both contracts and single premium contract and their associated prices.

Proposition 2 *Consider the dual contract is offered with P_H^d and P_L^d and the single premium or single basic contract is offered with P_H^s or P_L^s. We obtain both*

(a) offering dual contract dominates offering single premium contract and offering single basic contract and

(b) $P_H^d \geq P_H^s$ and $P_L^d \geq P_L^s$ if

$$\frac{(M_H + M_L - 2)^2}{4(1 - M_H)(1 - M_L)} < \frac{q_H}{q_L} \quad \text{and} \quad M_L < M_H.$$

When the dual contract is offered, the service provider expects to segment the market into two groups. The customers with high valuation v select the premium contract and pay a high price for the service. The provider then designs another basic service to attract low valuation customers. The former contributes the provider a high profit margin, and the latter helps the provider expand the market so that more customers are willing to purchase the service. These benefits become significant especially when the two services are highly differentiated owing to a large difference between the quality levels of the two services (i.e., high q_H/q_L). The beliefs of compensation ratio also play a role in the determination of the optimal service contract. The service provider expects that offering a dual contract is better than offering only a single premium or basic contract when either M_H and M_L is moderate (close to each other) or when both M_H and M_L are sufficiently large as long as $M_L < M_H$. In other words, the customers expect the compensation ratio to be similar regardless of which service contract they select or what compensation they can receive given that the penalty for unfulfillment is very high for each contract. The service provider intends to utilize both contracts simultaneously compared to adopting only one (premium or basic) contract, doing which successfully increases the prices of both contracts to increase the profit margin owing to the flexibility of both services. The provider can also set the prices of the two contracts at a high level to enhance profit.

Finally, the case where the service provider offers only one contract to customers is also investigated. The following proposition shows that the quality levels of the contracts offered and the corresponding beliefs of compensation ratio influence the service contract decisions.

Proposition 3 *Offering single premium contract dominates offering single basic contract if and only if $\frac{1-M_H}{1-M_L} < \frac{q_H}{q_L}$. Furthermore, the price for single premium contract is larger than that for single basic contract.*

M_H and M_L are the beliefs of compensation ratio when the requested units of resource cannot be fulfilled by the service provider from the customers' point of view. Thus, $1 - M_H$ and $1 - M_L$ can be regarded as the *fulfillment rate* of each service contract. The condition in the proposition reveals the relative relationship between the ratio of the fulfillment rate and the ratio of the quality levels for the two service contracts. For a given ratio of quality level, $\frac{q_H}{q_L}$, if the service provider is able to decrease $1 - M_H$ relative to $1 - M_L$, then offering single premium contract is appropriate; otherwise, the service provider is better off offering single basic contract.

In our basic model setting, we assume customers form the beliefs of compensation, M_H and M_L, given that customers are unable to know the capacity, T, and probabilities $g_{ij}, i = H, L, j = 1, 2$. Another way to model is to consider the customers may form rational expectation on the possible compensation. Due to analytical complexity of our basic model, we only focus on one special case where customers rationally expect that the service provider can always fulfill the required units of resource. This scenario exists if the service offered by the provider is quite reliable or customers believe the capacity of the service provider is significantly large, and hence, the customers do not expect any compensation. Under this scenario (i.e., when the unit of resource is

sufficiently large), we obtain that $P_H^d = P_H^s = \frac{q_H}{2}$ and $P_L^d = P_L^s = \frac{q_L}{2}$.[4] Also, the provider's profit when offering the dual contract is $\frac{q_H+q_L}{4}$, which is larger than the profit under single premium ($\frac{q_H}{4}$) and under single basic ($\frac{q_L}{4}$), due to market segmentation. In Section 7, we extend the model to a broad setting under which the beliefs can be endogenously determined. We analytically discuss how to obtain the beliefs based on different capacity levels and conduct numerical experiment to gain more insights.

4.2. When unit of resource is not sufficiently large

Now, the focus is moved to the case where T is not sufficiently large. If the service provider offers both premium and basic services, we can obtain the optimal prices for each case depending on resource T and based on the first-order conditions of the provider's profit function given that the profit function of the service provider is concave in both P_H and P_L. The optimal prices for each case are determined first without considering the associated constraint. To check whether the obtained prices are indeed optimal, the optimal prices are substituted into the constraint of each case. If the corresponding constraint is satisfied, then the prices are indeed optimal in that case; otherwise, the corner solution of the prices in that case requires consideration. Same logic also applies to the cases where a single premium or basic contract is offered.

Given that the analysis of the corner solution of prices in each case when the dual contract is offered is extremely complicated and messy, a numerical study is then conducted to further discuss the optimal prices for each case. The following proposition shows the optimal price for each case if the constraint of each case is satisfied when single contract is offered.

Proposition 4 *The optimal price when the service provider offers single premium contract and single basic contract is listed in Table 2.*

When offering single premium or basic contract, the optimal price is equal to the base price (i.e., P_{H3}^* or P_{L3}^*) plus a surplus (i.e., $P_{i1}^* - P_{i3}^*$ and $P_{i2}^* - P_{i3}^*$, $i \in \{L, H\}$). The determination of the surplus is based on the quality q_H and q_L, compensation ratios $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, probability of resource usage g_2, unit of resource T, resource redundancy factor δ, and the belief of compensation ratios M_H and M_L. It is not difficult to find that for everything else being equal an increase in δ leads to the reduction of the surplus when single premium contract is offered since a larger requirement of resource redundancy forces the service provider not to charge the price too high so as to avoid penalty cost if customer's request cannot be fulfilled. However, all the remaining factors positively enhance the increase in the surplus.

[4]One may simply set $M_H = M_L = 0$ in our basic model setting to derive the corresponding results.

5. Numerical study

Having obtained the analytical expressions, several numerical experiments are conducted to illustrate the structural properties of the problem and to gain more managerial insights regarding the optimal service contracts the service provider will select. The effects of the model characteristics on the profit of the service provider and associated optimal contracts offered to end customers are investigated.

Service provider's profit We first examine how different factors influence the service provider's profit. First, our results show that the service provider's profit increases in the belief of the compensation ratio for the premium service contract, M_H, and in the probability of requesting one unit of resource for premium service contract, g_{H1}. Given a high M_H, customers believe that they will receive a high compensation from the service provider if the unit of resource they request cannot be fulfilled for this high-end service. By taking advantage of this high compensation belief, the service provider charges a high price for the premium service contract and knows that this high price is still attractive to the high-end customers. As a result, high price strategy successfully gains more profit margin from customers, leading to a high profit by promoting the premium service contract. The effect of M_L on the profit of the service provider is similar for the basic contract as the provider can upward the price of low-end service accordingly. On the other hand, a large g_{H1} implies that a large portion of the customers request only one unit of resource when they purchase the premium contract, which mitigates the pressure on the service provider for not fulfilling enough resource to this group of customers (i.e., premium service customers). Therefore, the provider is able to gain more profit without worrying too many penalties paid to the customers who select premium service when g_{H1} is high. The same pattern can be observed for the effect of g_{L1} on the profit of the service provider as well, and thus, the associated figure is omitted.

Furthermore, the profit of the service provider is high when the quality level of the premium or basic service contract, q_H or q_L, is high. Note that high-quality level enhances the profit of the provider because the provider can increase the price by enhancing the quality of the service. Therefore, an increase in quality level benefits the service provider. Finally, the service provider receives less profit when each compensation ratio for the premium service contract, α_1 or α_2, increases. This pattern is particularly obvious when the unit of resource is scarce. When the resource is limited, the service provider cannot fully allocate available resource to each customer. Therefore, the penalty cost erodes the profit of the provider significantly when the compensation ratio is large. However, this effect diminishes when more resource is obtained. When the unit resource is sufficiently large, the service provider's profit does

Table 2 Optimal prices for single premium and basic contracts when unit of resource is not sufficiently large

Contract	*Case*	*Optimal price*
Premium	$T < (1+\delta)(X_{H1}+X_{H2})$	$P^*_{H1} = \frac{q_H \alpha_1 T}{2(1+\delta)(1-M_H)(1-\alpha_1-\alpha_2 g_2)} + \frac{q_H}{2(1-M_H)}$
	$(1+\delta)(X_{H1}+X_{H2}) \le T < (1+\delta)(X_{H1}+2X_{H2})$	$P^*_{H2} = \frac{q_H \alpha_2 T}{2(1+\delta)(1-M_H)(1-\alpha_2-\alpha_2 g_2)} + \frac{q_H}{2(1-M_H)}$
	$(1+\delta)(X_{H1}+2X_{H2}) \le T$	$P^*_{H3} = \frac{q_H}{2(1-M_H)}$
Basic	$T < X_{L1}+X_{L2}$	$P^*_{L1} = \frac{q_L \alpha_3 T}{2(1-M_L)(1-\alpha_3-\alpha_4 g_2)} + \frac{q_L}{2(1-M_L)}$
	$X_{L1}+X_{L2} \le T < X_{L1}+2X_{L2}$	$P^*_{L2} = \frac{q_L \alpha_4 T}{2(1-M_L)(1-\alpha_4-\alpha_4 g_2)} + \frac{q_L}{2(1-M_L)}$
	$X_{L1}+2X_{L2} \le T$	$P^*_{L3} = \frac{q_L}{2(1-M_L)}$

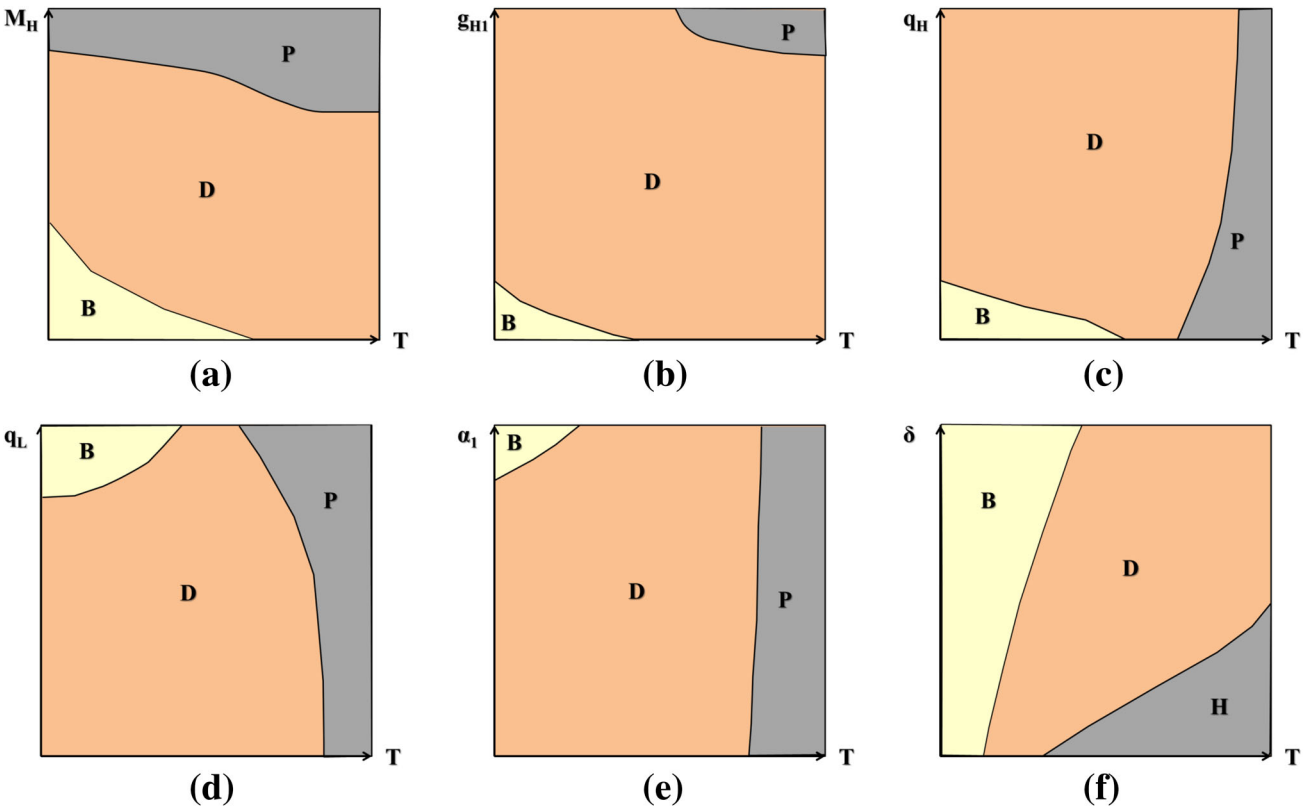

Figure 1 Effects of model characteristics on the optimal service contracts of the service provider. Here, *B* basic contract, *P* premium contract, and *D* dual contract.

not depend on the compensation ratio, α_1 or α_2, given that all the requested resources can be fully satisfied.

Optimal service contracts We concentrate on the optimal contract, namely dual contract or single contract, that the service provider will offer. We summarize the results in Figure 1[5]. From Figure 1a, b, we observe that the service provider tends to offer single basic contract when both M_H and g_{H1} are low. This outcome is more significant when the resource is limited. Notice that a low M_H reduces the willingness of the customers to purchase the premium contract since customers do not expect to receive a high portion of compensation if the requested resource cannot be fulfilled. Also, a low g_{H1} represents a high probability of requesting two units of resource if the customers select the premium contracts given that $g_{H1} = 1 - g_{H2}$. Both variables influence the profit of the provider from the premium service as the former reduce the possibility of receiving a higher profit margin and the latter intensifies the burden of the service provider if the resource is limited. Given the drawbacks of offering the premium service contract, offering only a single basic contract would be better for the service provider. However, when M_H and/or g_{H1} is high, the abovementioned effect reverses as the service

[5]Unless otherwise specified, we use $g_{H1} = 0.25, g_{L1} = 0.7, g_{H2} = 0.75, g_{L2} = 0.3, M_H = 0.07, M_L = 0.03, q_H = 0.9, q_L = 0.7, \alpha_1 = 0.1, \alpha_2 = 0.06, \alpha_3 = 0.03, \alpha_4 = 0.02,$ and $\delta = 0.5$.

provider offers a single premium contract targeting customers with high willingness to pay and at the same time without worrying too much the penalties paid to these customers. This trend is more apparent when the resource is sufficient. When both M_H or g_{H1} are modest, the provider needs to balance these two driving forces at the same time and the dual contract is then offered. By offering the dual contract, the service provider adopts two services to segment the end market into two groups and adjust the prices to do better price discrimination.

Analysis of how the quality levels influence the optimal contracts (Figure 1c, d) indicates that when q_H and q_L are close (i.e., low q_H or high q_L), two contracts are similar from the customer's perspective. Thus, price differentiation does not benefit profit. Hence, a single basic contract is offered, especially when the resource is limited. When the two quality levels differ, the effect of market segmentation improves the profit of the provider. The service provider has the incentive to segment the market by offering two different service contracts and a dual contract is offered. When the resource is sufficient and the quality of the premium contract is advanced, the effect of market segmentation is dominated by the large margin provided by the premium contract without paying a large amount of penalty cost. Hence, single premium contract is the optimal service contract.

Finally, Figure 1e, f shows the effects of compensation ratios α_1 and redundancy factor δ on the optimal contracts. Note that the effect of the compensation ratio is contrary to that of M_H or g_{H1}. When the compensation ratio α_1 increases, the optimal outcome switches from single premium contract to dual contract and then to single basic contract because when α_1 increases, the service provider would be penalized more if the resource requested by the premium contract customers is not fulfilled. The effect becomes severe when the resource is limited since there is higher probability that the service provider is penalized for not being able to satisfy the requested resource. Therefore, single basic contract dominates. When the compensation ratio is reduced, the service provider offers a single premium contract to enhance the profit margin. Also, sufficient resource helps the provider achieve this goal by avoiding penalty costs. A dual contract simply balances the two effects and is offered when the ratios are modest. An increase in the redundancy factor δ also disallows the provider to adopt single premium contract since more units of resource are needed for premium customers.

6. Network externality

In this section, the current model is extended to investigate the effect of network externality of cloud computing service. In order to analyze this effect more generally, we consider both positive and negative network externality. Define β as the factor of network externality and $\beta > 0$ (<0) represents the case where an increase in usage of cloud service leads to a decrease (an increase) for other users. We also define $X = X_{H1} + X_{H2} + X_{L1} + X_{L2}$ as the total number of customers who sign the contracts with the provider and use the cloud service where each $X_{ij}, i = H, L, j = 1, 2$ is defined earlier. A customer's willingness to pay, defined as $v - \beta X$, is affected by the number of the customers who sign the contract and use the cloud service. A positive (negative) β shows a larger number of users discourage (encourage) people to use the cloud service and, hence, reduce (raise) the willingness to pay of a customer. A customer's net utility of selecting a service contract is $V_i = (v - \beta X)q_i - P_i + M_i P_i$ where i is H or L. To analyze the model, we follow the same logic by obtaining two thresholds: $\underline{v}$ and $\bar{v}$, which are described below:

$$\begin{aligned} \underline{v} &= \frac{(1-M_L)P_L}{q_L} + \beta X = \frac{(1-M_L)P_L}{q_L} + \beta(1-\underline{v}), \\ \bar{v} &= \frac{(1-M_H)P_H - (1-M_L)P_L}{q_H - q_L} \\ &\quad + \beta X = \frac{(1-M_H)P_H - (1-M_L)P_L}{q_H - q_L} + \beta(1-\underline{v}). \end{aligned} \tag{9}$$

After algebra, we obtain

$$\begin{aligned} \underline{v} &= \frac{(1-M_L)P_L}{(1+\beta)q_L} + \frac{\beta}{1+\beta}, \quad \bar{v} = \frac{(1-M_H)P_H - (1-M_L)P_L}{q_H - q_L} \\ &\quad + \beta\left(1 - \frac{(1-M_L)P_L}{(1+\beta)q_L} - \frac{\beta}{1+\beta}\right). \end{aligned} \tag{10}$$

Discussion We analyze the model and find the optimal solutions under each type of contract by taking into account network externality. We mainly focus on the effect of network externality β on the optimal prices and profits and the associated optimal contract design. Figure 2 illustrates the optimal service contract in terms of effect of network externality and resource. Consider first the effect of negative network externality ($\beta > 0$). When the effect of negative network externality increases, the service provider is more likely to choose single premium contract, especially when the unit of resource T is large. The logic behind this observation is as follows. When offering dual contract, the service provider can gain additional profit due to market segmentation. However, this benefit of adopting dual contract enhances the market share but at the same time reduces the willingness to pay of each customer due to negative network externality. Instead of offering dual contract, the service provider focuses on high willingness to pay customers by offering single premium contract, doing which can increase the price and also the profit margin. This high price strategy can successfully weaken the impact from the negative externality. Therefore, offering single premium contract is better when the effect of negative network externality is apparent. On the other hand, when the effect of negative network externality decreases, the impact of market share on the contract design is not significant as the unit of resource plays an important role on the contract

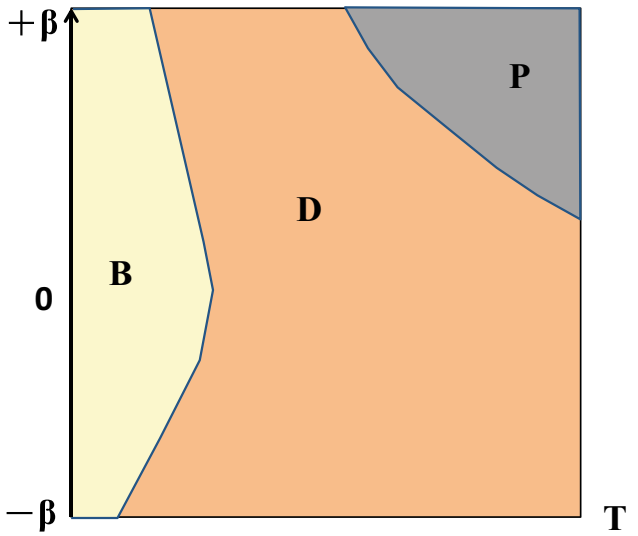

Figure 2 Effects of network externality and unit of resource on the optimal contract. Here, B basic contract, P premium contract, and D dual contract. We use $g_{H1} = 0.25, g_{L1} = 0.7, g_{H2} = 0.75, g_{L2} = 0.3, M_H = 0.07, M_L = 0.03, q_H = 0.9, q_L = 0.7, \alpha_1 = 0.1, \alpha_2 = 0.06, \alpha_3 = 0.03, \alpha_4 = 0.02$, and $\delta = 0.5$.

design. We observe from Figure 2 that with limited unit of resource, the service provider tends to choose single basic contract to avoid the constraint of resource redundancy. When both negative network externality and unit of resource are modest, dual contract is optimal.

When there exists positive network externality (i.e., $\beta < 0$), our results show that the service provider offers dual contract when the effect is significant and/or the unit of resource T is large and the single basic contract otherwise. In particular, we show that single premium contract is not offered when $\beta < 0$. Note that negative β induces the service provider to offer dual contract to attract more customers. However, the consideration of resource redundancy discourages the provider to offer single premium contract. Hence, when β is small (positive externality is significant) and/or T is large, the offer of dual contract benefits the provider from not only market segmentation but also customers' high willingness to pay due to the positive externality effect. When β increases and is close to zero (positive effect is minor), the result is mainly led by the unit of resource T as aforementioned discussion under the case $\beta > 0$. Therefore, we observe similar results.

7. Extension

In this section, we consider two extensions: (1) variant compensation ratio and (2) beliefs are endogenously determined.

7.1. Variant compensation ratio

In our model setting, we assume customers who sign the premium contract always possess higher priority and hence, $\alpha_4 \le \alpha_3 \le \alpha_2 \le \alpha_1$. One may be interested in the case in which the customers signing the basic contract have a higher priority to obtain the first unit of resource than the ones who sign the premium contract for the second unit, i.e., $\alpha_4 \le \alpha_2 \le \alpha_3 \le \alpha_1$. Under this situation, fulfilling the first unit of resource to customers who sign the premium contract has the highest priority followed by offering the first unit of resource to customers who sign the basic contract. The remaining resource will be assigned to customers who request the second unit for the same logic. Under the new allocation rule, five different cases are also obtained, while the service provider offers both premium and basic contracts. The analysis of Cases 1, 4, and 5 remains the same as in Section 3.1. The original Cases 2 and 3 are changed to Cases 2A and 3A below:

Case 2A: When T is relatively small: $(1+\delta)(X_{H1}+X_{H2}) \le T < (1+\delta)(X_{H1}+X_{H2}) + X_{L1} + X_{L2}$

When $(1+\delta)(X_{H1}+X_{H2}) \le T$, all premium contract customers can obtain their first unit of resource. The rest of the resource will be allocated to fulfill the first unit of the customers with basic contract. Therefore, the service provider chooses P_H and P_L to maximize the profit π_B^2 subject to $(1+\delta)(X_{H1}+X_{H2}) \le T < (1+\delta)(X_{H1}+X_{H2}) + X_{L1} + X_{L2}$, where

$$\begin{aligned}\pi_B^2 = {} & (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L \\ & - ((1+\delta)(X_{H1}+X_{H2}) + X_{L1} + X_{L2} - T)\alpha_3 P_L \\ & - X_{H2}\alpha_2 P_H - X_{L2}\alpha_4 P_L.\end{aligned} \tag{11}$$

Case 3A: When T is modest: $(1+\delta)(X_{H1}+X_{H2}) + X_{L1} + X_{L2} \le T < (1+\delta)(X_{H1}+2X_{H2}) + X_{L1} + X_{L2}$

In this case, the service provider is able to fulfill the first unit of resource for all customers and starts to assign the additional resource for the premium contract customers requesting two units. Therefore, the service provider sets prices P_H and P_L to maximize the profit

$$\begin{aligned}\pi_B^3 = {} & (X_{H1}+X_{H2})P_H + (X_{L1}+X_{L2})P_L \\ & - \left(\frac{(1+\delta)(X_{H1}+2X_{H2}) + X_{L1} + X_{L2} - T}{1+\delta}\right) \\ & \alpha_2 P_H - X_{L2}\alpha_4 P_L\end{aligned}$$

with the constraint

$$\begin{aligned}(1+\delta)(X_{H1}+X_{H2}) + X_{L1} + X_{L2} \le T < {} & (1+\delta)(X_{H1}+2X_{H2}) \\ & + X_{L1} + X_{L2}.\end{aligned} \tag{12}$$

Similarly, given the unit of resource, T, the service provider solves the five cases separately to determine optimal prices P_H and P_L in each case. The service provider then compares the profit that can be obtained from the cases to determine the optimal prices to be adopted.

Based on the results, we find that the patterns of the equilibrium strategies of the service provider are similar to the original setting, i.e., $\alpha_2 \ge \alpha_3$. Under variant compensation ratio

where $\alpha_2 \leq \alpha_3$, however, basic contract customers are more valuable to the provider as the provider needs to fulfill those who request one unit of resource under basic contract. Under this setting, the effect of price discrimination is reduced. Hence, everything being equal, we find that the service provider is more likely to adopt single contract (basic or premium) rather than the dual contract.

7.2. Beliefs are endogenously determined

In our basic setting, we assume the beliefs are exogenously given and obtain equilibrium solutions. In this subsection, we consider the case where the beliefs, M_H and M_L, are endogenously determined given the customers and the service provider possess the same information, i.e., the customers know the probability $g_{ij}, i = \{L, H\}, j = 1, 2$ and resource T. To obtain the equilibrium M_H and M_L, we reconsider five cases in Section 3.1. Consider first the case where T is sufficiently large (i.e., Case 5 of Section 3.1), the service provider is able to fulfill all the requested units, and hence, we can obtain that $M_H = M_L = 0$. Furthermore, for Case 4 where T is relatively large, all the requested units from premium contract customers are fulfilled. Some of the basic contract customers who request two units of resource will receive compensation. Therefore, we can obtain $M_H = 0$ and $M_L = \frac{(X_{H1}+2X_{H2})(1+\delta)+X_{L1}+2X_{L2}-T}{X_{L1}+X_{L2}}\alpha_4$, where $\frac{(X_{H1}+2X_{H2})(1+\delta)+X_{L1}+2X_{L2}-T}{X_{L1}+X_{L2}}$ represents the probability of basic contract customers who cannot obtain the second unit of resource. We summarize all the results in Tables 3 and 4.

Note that under Section 4.1 where T is sufficiently large, we show how the optimal prices are influenced by model characteristics. When the beliefs are endogenously determined, all the results continue to hold by setting $M_H = M_L = 0$. Also, the result in Proposition 4 can be adjusted by substituting M_H and M_L derived by Table 4 into the prices in Table 2. In addition, we are mainly interested in how the optimal service contracts offered by the provider are affected if M_H and M_L are endogenously determined. We obtain that the service provider adopts single basic contract when the capacity T (or g_{H1}) is low and switches to the dual contract and then single premium contract accordingly as T (or g_{H1}) increases. On the other hand, the effect of q_L leads to the opposite result as the service provider is more likely to switch from the dual contract to the single basic contract as q_L increases. All the aforementioned results basically follow the similar patterns as in Figure 1.

8. Conclusion

In a fast-moving network access environment, cloud computing service providers can supply a variety of resources such as file storage, computational processing, and software applications to contracted customers. In addition to price, the service contracts offered by the service providers also specify the SLA with the corresponding penalty charge once service availability cannot be guaranteed by certain conditions. Creating different contracts with varied SLAs and penalties allows the service provider to conduct price differentiation to better categorize the market and

Table 3 M_H and M_L for dual contract

Contract	*Case (constraint)*	M_H *and* M_L
Dual	Case 1	$M_H = \frac{(1+\delta)(X_{H1}+X_{H2})-T}{(1+\delta)(X_{H1}+X_{H2})}\alpha_1 + \frac{X_{H2}}{X_{H1}+X_{H2}}\alpha_2$, $M_L = \alpha_3 + \frac{X_{L2}}{X_{L1}+X_{L2}}\alpha_4$
	Case 2	$M_H = \frac{(1+\delta)(X_{H1}+2X_{H2})-T}{(1+\delta)(X_{H1}+X_{H2})}\alpha_2$, $M_L = \alpha_3 + \frac{X_{L2}}{X_{L1}+X_{L2}}\alpha_4$
	Case 3	$M_H = 0$, $M_L = \frac{(1+\delta)(X_{H1}+2X_{H2})+X_{L1}+X_{L2}-T}{X_{L1}+X_{L2}}\alpha_3 + \frac{X_{L2}}{X_{L1}+X_{L2}}\alpha_4$
	Case 4	$M_H = 0$, $M_L = \frac{(X_{H1}+2X_{H2})(1+\delta)+X_{L1}+2X_{L2}-T}{X_{L1}+X_{L2}}\alpha_4$
	Case 5	$M_H = 0$, $M_L = 0$

Table 4 M_H and M_L for single premium and single basic contracts

Contract	*Case (constraint)*	M_H *and* M_L
Premium	$T < (1+\delta)(X_{H1}+X_{H2})$	$M_H = \frac{(1+\delta)(X_{H1}+X_{H2})-T}{(1+\delta)(X_{H1}+X_{H2})}\alpha_1 + \frac{X_{H2}}{X_{H1}+X_{H2}}\alpha_2$
	$(1+\delta)(X_{H1}+X_{H2}) \leq T < (1+\delta)(X_{H1}+2X_{H2})$	$M_H = \frac{(1+\delta)(X_{H1}+2X_{H2})-T}{(1+\delta)(X_{H1}+X_{H2})}\alpha_2$
	$(1+\delta)(X_{H1}+2X_{H2}) \leq T$	$M_H = 0$
Basic	$T < X_{L1}+X_{L2}$	$M_L = \frac{X_{L1}+X_{L2}-T}{X_{L1}+X_{L2}}\alpha_3 + \frac{X_{L2}}{X_{L1}+X_{L2}}\alpha_4$
	$X_{L1}+X_{L2} \leq T < X_{L1}+2X_{L2}$	$M_L = \frac{X_{L1}+2X_{L2}-T}{X_{L1}+X_{L2}}\alpha_4$
	$X_{L1}+2X_{L2} \leq T$	$M_L = 0$

further optimize profit. A model was constructed in this research to compare profit performance in a dual contract and single contract market with the consideration of resource redundancy. The optimal prices for each contract were derived based on the rule of resource allocation and depending on the different resource capacities. The results show that when the resource is sufficiently large, the optimal prices and the associated optimal contract to be offered are sensitive to whether the two services are differentiated and how the customers perceive the ratio for compensation. When the two services are differentiated and both beliefs of compensation ratios are modest, offering the dual contract is the optimal; the prices for both contracts are increased to enhance the provider's profit.

Numerical experiments were also conducted to gain more managerial insights. The numerical examples show that the service provider is able to increase profit by enhancing the customers' beliefs of compensation ratio and the quality of services, thereby inducing customers with low willingness to pay to sign the premium contract. The belief of compensation ratio and the quality levels of the services also influence the optimal service contracts offered by the service provider. Another key factor affecting the optimal service contract is the resource acquired by the service provider. A sufficient resource eases the negative effect of penalty costs and allows the service provider to design contracts with flexibility. Several extensions are also considered to further investigate the equilibrium strategies.

Acknowledgements The authors thank the editor and the referees for helpful comments and suggestions. This work was supported by NSC grants NSC 100-2219-E-002-026, NSC 101-2219-E-002-017, and MOST 103-2221-E-002 -219-MY2 in Taiwan.

References

Aloi G, Musmanno R, Pace P and Pisacane O (2012). A wise cost-effective supplying bandwidth policy for multilayer wireless cognitive networks. *Computers and Operations Research* **39**(11):2836–2847.

An B, Lesser V, Irwin D and Zink M (2010). Automated negotiation with decommitment for dynamic resource allocation in cloud computing. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, vol. 1, pp. 981–988.

Anandasivam A and Premm M (2009). Bid price control and dynamic pricing in clouds. In: *17th European Conference on Information Systems, ECIS 2009 Proceedings*, 284.

Ardagna D, Trubian M and Zhang L (2007). SLA based resource allocation policies in autonomic environments. *Journal of Parallel and Distributed Computing* **67**(3):259–270.

Beloglazov A, Abawajy J and Buyya R (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems-the International Journal of Grid Computing and Escience* **28**(5):755–768.

Bhargava HK and Sun D (2008). Pricing under quality of service uncertainty: Market segmentation via statistical qos guarantees. *European Journal of Operational Research* **191**(3):1189–1204.

Bitran GR and Mondschein SV (1997). Periodic pricing of seasonal products in retailing. *Management Science* **43**(1):64–79.

Candogan O, Bimpikis K and Ozdaglar A (2012) Optimal pricing in networks with externalities. *Operations Research* **60**(4): 883–905.

Fawaz W, Daheb B, Audouin O, Du-Pond M and Pujolle G (2004). Service level agreement and provisioning in optical networks. *Communications Magazine, IEEE* **42**(1):36–43.

Fishburn PC and Odlyzko AM (2000). Dynamic behavior of differential pricing and quality of service options for the internet. *Decision Support Systems* **28**(1–2):123–136.

Fulp EW and Reeves DS (2004). Bandwidth provisioning and pricing for networks with multiple classes of service. *Computer Networks-the International Journal of Computer and Telecommunications Networking* **46**(1):41–52.

Ganesh A, Laevens K, Steinberg R (2007). Congestion pricing and noncooperative games in communication networks. *Operations Research* **55**(3):430–438.

Guan Y, Yang W, Owen H and Blough DA (2008). A pricing approach for bandwidth allocation in differentiated service networks. *Computers and Operations Research* **35**(12):3769–3786.

Hosanagar K, Krishnan R, Chuang J and Choudnary V (2005). Pricing and resource allocation in caching services with multiple levels of quality of service. *Management Science* **51**(12):1844–1859.

Jain S and Kannan PK (2002). Pricing of information products on online servers: Issues, models, and analysis. *Management Science* **48**(9):1123–1142.

Jukic B, Simon R and Chang WS (2004). Congestion based resource sharing in multi-service networks. *Decision Support Systems* **37**(3):397–413.

Kasap N, Aytug H and Erenguc SS (2007). Provider selection and task allocation issues in networks with different qos levels and all you can send pricing. *Decision Support Systems* **43**(2): 375–389.

Kate ML and Shaprio C (1994). Systems competition and network effects. *The Journal of Economic Perspectives* **8**(2):93–115.

Keskin T and Taskin N (2014). A pricing model for cloud computing service. In: *System Sciences (HICSS), 2014 47th Hawaii International Conference*, pp. 699–707.

Lin WY, Lin GY and Wei HY (2010). Dynamic auction mechanism for cloud resource allocation. In: *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference*, pp. 591–592.

Liu T, Methapatara C and Wynter L (2010). Revenue management model for on-demand it services. *European Journal of Operational Research* **207**(1):401–408.

Mihailescu M and Teo YM (2010). Dynamic resource pricing on federated clouds. In: *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference*, pp 513–517.

Ou J, Parlar M and Sharafali M (2006). A differentiated service scheme to optimize website revenues. *Journal of the Operational Research Society* **57**(11):1323–1340.

Rouskas AN, Kikilis AA and Ratsiatos SS (2008). A game theoretical formulation of integrated admission control and pricing in wireless networks. *European Journal of Operational Research* **191**(3): 1175–1188.

Shapiro C and Varian HR (1998). *Information Rules - A Strategic Guide to the Network Economy*. Cambridge, USA: Harvard Business School Press.

Sundararajan A (2004). Nonlinear pricing of information goods. *Management Science* **50**(12):1660–1673.

Teng F and Magoules F (2010). Resource pricing and equilibrium allocation policy in cloud computing. In: *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference*, pp. 195–202.

Thomas P, Teneketzis D and Mackie-Mason JK (2002). A market-based approach to optimal resource allocation in integrated-services connection-oriented networks. *Operations Research* **50**(4):603–616.

Wei GY, Vasilakos AV, Zheng Y and Xiong NX (2010). A game-theoretic method of fair resource allocation for cloud computing services. *Journal of Supercomputing* **54**(2):252–269.

Wu L and Buyya R (2010). Service level agreement (SLA) in utility computing systems. arXiv:1010.2881.

Zhang Z, Dey D and Tan Y (2008). Price and qos competition in data communication services. *European Journal of Operational Research* **187**(3):871–886.

Zhang Z, Tan Y and Dey D (2009). Price competition with service level guarantee in web services. *Decision Support Systems*, **47**(2):93–104.

Received 26 May 2015;
accepted 18 October 2016

Electronic supplementary material The online version of this article (doi:10.1057/s41274-016-0141-z) contains supplementary material, which is available to authorized users.