Lee *Journal of Cloud Computing: Advances, Systems and Applications* (2019) 8:15
https://doi.org/10.1186/s13677-019-0140-0

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

An optimization approach to capacity evaluation and investment decision of hybrid cloud: a corporate customer's perspective

In Lee

Abstract

While the rapid growth of cloud computing is driven by the surge of big data, the Internet of Things, and social media applications, an evaluation and investment decision for cloud computing has been challenging for corporate managers due to a lack of proper decision models. This paper attempts to identify critical variables for making a cloud capacity decision from a corporate customer's perspective and develops a base mathematical model to aid in a hybrid cloud investment decision under probabilistic computing demands. The identification of the critical variables provides a means by which a corporate customer can effectively evaluate various cloud capacity investment opportunities. Critical variables included in this model are an actual computing demand, the amount of private cloud capacity purchased, the purchase cost of the private cloud capacity, the price of the public cloud, and the default downtime loss/penalty cost. Extending the base model developed, this paper also takes into consideration the interoperability cost incurred in cloud bursting to the public cloud and derives the optimal investment. The interoperable cloud systems require time and investment by the users and/or cloud providers and there exists a diminishing return on the investment. Hence, the relationship between the interoperable cloud investment and return on investment is also investigated.

Keywords: Hybrid cloud, Interoperability, Public cloud, Private cloud, Decision model, Optimization, ROI

Introduction

The evolution of the Internet, storage technologies, service-oriented architecture, and grid computing has led to the development of cloud computing. Cloud computing has emerged as a disruptive innovation offering a variety of computing services and resources to individual users and corporate customers. Cloud infrastructure that was traditionally limited to single provider data centers is now evolving to the use of infrastructure from multiple providers [1]. A wide variety of deployment models, service models, and pricing schemes are flexibly integrated with each other to help enterprises transform the way they conduct business and meet their idiosyncratic computing needs.

Cloud computing complements traditional client-server computing to support the computing needs of businesses with a range of benefits such as lower IT expenditures, resource pooling, single source application updates, and scalability. Many cloud computing providers such as Google, Microsoft, IBM, and Amazon have been heavily investing in cloud technology and are leading various cloud services markets [2]. However, the increasing dissemination of cloud services and the growing number of cloud service providers (CSPs) have resulted in uncertainty for corporate customers in adopting cloud services [3]. Opara-Martins, Sahandi, and Tian [4] find that the most cited reasons for adopting cloud computing include better scalability of IT resources (45.9%), collaboration (40.5%), cost savings (39.6%), and increased flexibility (36.9%).

The most cited challenge among corporate customers is managing costs, but they underestimate the amount of wasted cloud expense [5]. Respondents estimate 27% waste, while RightScale [5] has measured actual waste at 35%. 84% of the respondents with more than 1000 employees are

Correspondence: I-Lee@wiu.edu
Cecil P. McDonough Endowed Professor in Business, School of Computer Sciences, Western Illinois University, Macomb, IL 61455, USA

using a multi-cloud strategy and 58% are using the hybrid cloud. Cloud cost saving is the top priority across all corporate customers (64%). Therefore, this study cannot emphasize enough the importance of the solid justification of the economic capacity evaluation and management of cloud computing as a capacity evaluation and investment decision.

Capacity planning is a challenging task when there is an unpredictable, fluctuating computing demand with many peaks and troughs [6]. Without a solid evaluation model, estimating the tradeoff between the benefits and costs incurred in order to cover peak computing demand is challenging. Therefore, overcapacity or under-capacity is a common phenomenon in the investment of cloud capacity. Overcapacity puts companies at a cost disadvantage due to a low utilization of cloud resources. On the other hand, under-capacity puts them at a strategic disadvantage due to customer/user dissatisfaction, high penalty costs, and potential sales loss.

The hybrid cloud enjoys the benefits of both private and public clouds, but the combination of the two has introduced challenging issues such as data security, performance, and cost optimization [7]. According to RightScale's 2019 State of the Cloud Report [5], the hybrid cloud is the most preferred cloud for corporate cloud capacity investment. The hybrid cloud is flexible enough to handle the spike of computing demand while at the same time reducing computing costs. The hybrid cloud is a cloud environment in which an organization owns and manages their internal private cloud and uses a public cloud or a community cloud externally when necessary. In the hybrid cloud, typically, non-critical computing resources are outsourced to the public cloud, while mission-critical applications and data are kept in the private cloud under the strict control of the organization. Researchers used the MapReduce paradigm to split a data-intensive workload into mapping tasks sorted by the sensitivity of the data, with the most sensitive data being processed at on-premise servers and the least sensitive processed in a public cloud [8]. Many leading enterprises such as Uber, GM, and JP Morgan are adopters of the hybrid cloud. For example, Uber relies on a hybrid cloud infrastructure that combines the use of public cloud services with standardized on-premise server racks in its datacenters [9].

Whether to utilize cloud computing or internal IT resources for business purposes is a critical decision for organizations and decision makers [10]. For companies, the cloud evaluation and investment decision to minimize the total computing costs requires a basic understanding of the relationships between the cost of computing resources, a penalty for computer downtime, and a probability distribution of the computing demand. In light of the lack of studies in the cloud investment area and the corporate trend of migrating to the hybrid cloud, this paper proposes an optimization approach to cloud evaluation and an investment decision of the hybrid cloud for corporate decision makers. The remainder of this paper includes a literature review on cloud capacity investment decisions, the base decision model, an illustration of the model operation, a model extension with an interoperability cost for cloud bursting, and return on investment (ROI) management.

Literature review

While exact modeling of cloud centers is not feasible due to the nature of cloud centers and the diversity of user requests [11], workload modeling has been used to increase the understanding of typical cloud workload patterns and has led to more informed resource management decisions [12]. Workload modeling and characterization is especially challenging when applied in a highly dynamic environment for reasons such as heterogeneous hardware in a single data center and complex workloads composed of a wide variety of applications with different characteristics and user profiles [13]. In a large-scale system faced by time varying and regionally distributed demands for various resources, there is a tradeoff between optimizing the resource placement to meet its demand and minimizing the number of added and removed resources to the placement. Novel analytic techniques utilizing graph theory methodologies are proposed that overcome this difficulty and yield very efficient dynamic placements with bounded repositioning costs [14].

A number of studies on models of computing resources assume the exponential distribution or the Weibull distribution of computing requests and conduct mathematical analyses of system performance [11, 15, 16]. For example, a cloud center is modeled as a classic open network with a single arrival from which the distribution of response time is obtained, assuming that both interarrival and service times are exponential [17]. A recent study of Patch and Taimre [18] also assumes that tasks require an exponentially distributed service time for transient provisioning and performance evaluation of cloud computing platforms.

Wolski and Brevik [19] conduct a simulation of workload patterns with real data and compare the efficacy of the lognormal distribution, the 3-phase hyper exponential distribution, and the pareto distribution and find that the use of the 3-phase hyper exponential distribution is the most appropriate based on the Kolmolgrov-Smirnov statistic which has been widely used to compute the goodness-of-fit between the observed data and a distribution fit to it. With the goal of providing a repository where practitioners and researchers can exchange grid workload traces, a research team from the Delft University of Technology maintains an archive that contains an extensive set of grid workload data as well as the results of the Kolmolgrov-Smirnov (KS) tests [20].

The Kolmolgrov-Smirnov static of the various workloads show that overall, the Weibull, the lognormal, and the gamma distributions have a better fit than others. While a workload analysis can be used for scheduler design, it is also useful for capacity planning.

The efficient and accurate assessment of cloud-based infrastructure is essential for guaranteeing both business continuity and uninterrupted services for computing jobs [21]. However, efficient resource management of the cloud infrastructures is challenging [22]. In the capacity planning and management area, most studies focus on micro-level scheduling such as dynamic resource allocation and prioritization of computing jobs. Widely used resource management methods such as Amazon Web Services (AWS) Auto Scaling and Windows Azure Autoscaling Application Block, are reactive. While these reactive approaches are an effective way to improve system availability, optimize costs, and reduce latency, it exhibits a mismatch between resource demands and service provisions, which could lead to under or over provisioning [23]. Hence, predictive resource scaling approaches have been proposed to overcome the limitations of the reactive approaches most often used [24–26].

Misra and Monda [27] point out that a challenge arises when companies with an existing data center make a decision on whether cloud computing would be helpful for them or if they should stick to their own in-house infrastructure for the expansion and consolidation of the existing data centers. They analyze the cost side aspect with a simulation model that not only helps assess the suitability of the cloud computing but also measures its profitability. Recently, Patch and Taimre [18] demonstrate that taking into account each of the characteristics of fluctuating user demand of cloud services can result in a substantial reduction of losses. Their transient provisioning framework allows for a wide variety of system behaviors to be modeled in the presence of various model characteristics through the use of matrix analytic methods to minimize the revenue loss over a finite time horizon.

de Assunção, di Costanzo, and Buyya [28] investigate the benefits that organizations can reap by using a public cloud to augment their local computing capacity. They evaluate six scheduling strategies to understand how these strategies achieve a balance between performance and usage cost, and how much they improve response time. Gmach, Rolia, Cherkasova, and Kemper [29] describe a capacity management process for resource pools. They use a trace-based approach which predicts future per-server required capacity and achieve a 35% reduction in processor usage as compared to the benchmark best practice.

The hybrid cloud can potentially reduce the financial burden of overcapacity investment and technological risks related to a full ownership of computing resources and can allow companies to operate at a cost-optimal scale and scope under demand uncertainty [30]. The hybrid cloud allows users to scale computing requirements beyond the private cloud and into the public cloud - a capability called cloud bursting in which an application runs in its own private resources for the majority of its computing and bursts into the public cloud when its private resources are insufficient for the peaks in computing demand [31]. For example, a popular and cost-effective way to deal with the temporary computational demand of big data analytics is hybrid cloud bursting that leases temporary off-premise cloud resources to boost the overall capacity during peak utilization [32]. Academic research into the hybrid cloud has focused on the middleware and abstraction layers for creating, managing, and using the hybrid cloud [8]. Commercial support for the hybrid cloud is growing in response to the increasing demand for the hybrid cloud. Container technologies will improve portability between different cloud providers.

While potential benefits of the hybrid cloud arise in the presence of variable demand for many real-world computing workloads, additional costs related to hybrid cloud management, data transfer, and development complexity must be taken into account [33]. Bandwidth, latency, location of data, and communication performance need to be considered for integrating a public cloud with a private cloud [34]. The interoperability and portability issues between the public cloud and the on-premise private cloud also continue to be barriers to its adoption. These issues arise when cloud providers develop services with non-compatible proprietary technologies [4]. While various standardized solutions have been developed for diverse cloud computing services [35], cloud providers often develop their own proprietary services as a way to lock in clients, differentiate their services, and achieve a market monopoly in the early stages of innovation. A lack of standardization poses challenges to cloud users who need to integrate diverse cloud services obtained from multiple cloud providers [36] and perform cloud bursting in the hybrid cloud environment.

While the above-mentioned studies investigate micro-level scheduling schemes for cloud resources, few studies have analyzed how cloud capacity for the public, private, and hybrid cloud interacts with the service level and prices in a given decision horizon. Furthermore, most previous studies did not attempt to develop any closed form solution for an optimal cloud capacity decision.

Cloud computing is in an expansion stage of technological diffusion and is projected to grow more rapidly with the advances in big data, artificial intelligence, and the Internet of Things. Given the growing interest in the cloud capacity decision by managers, researchers need to develop cost-benefit evaluation models and tools that will help managers make a judicious cloud capacity decision. The development of a quantifiable cost-benefit decision model is of vital importance to avoid intuition-based gut-

feeling investment decisions. In an attempt to fill a gap in the current studies, the next section provides a normative model for the cloud capacity decision.

A capacity evaluation and investment decision model of hybrid cloud

This section proposes a capacity evaluation and investment decision model of the hybrid cloud and examines the relationships between model parameters and cloud investments. This model derives the optimal capacity decision of the private cloud and the public cloud to minimize the total computing costs. The optimal decision is based on the cost structure and a probability density function for computing demand in a given period. This model reflects customers' perspectives and utilizes a model presented by Henneberger [37] which extends the classic newsvendor problem and derives a critical fractile formula using an inverse cumulative distribution function of computing demand. This newsvendor problem has been widely used in inventory management. For example, Fan et al. [38] apply the newsvendor model to analyze the reduction of inventory shrinkage with the use of RFID. While Henneberger's model did not present a closed-form solution, the proposed model derives a closed-form solution for a capacity evaluation and decision problem with an exponential probability distribution of computing demand. Later, the base model is extended to find the optimal solution for two investment decision variables simultaneously: (1) the optimal cloud capacity for the private cloud and (2) the optimal investment in the interoperability enhancement for cloud bursting in the hybrid cloud environment.

A hypothetical cloud capacity evaluation and investment decision problem

A company needs to decide how much private cloud they need to invest in and how much public cloud they need to use to minimize the total computing costs in a given decision horizon. While the company can choose to use a private cloud only, a public cloud only, or a hybrid cloud, they realize that the hybrid cloud has a cost advantage when a company has a wide range of peaks and troughs in computing demand. The company can purchase a private cloud capacity in its own virtual machine (VM) environment, and can use the pay-as-you-go public cloud for peak-time computing demands. The proposed model helps find an optimal mix of the private and public cloud to minimize the total computing costs. The following is the nomenclature used throughout this paper.

Nomenclature

x: an actual computing demand occurring in each time unit with an exponential probability distribution function

$\lambda e^{-\lambda x}$: an exponential probability distribution function for computing demand

$1\text{-}e^{-\lambda x}$: a cumulative exponential distribution function for computing demand

c: units of private cloud capacity purchased at the beginning of a decision period

k: one-time purchase cost per unit of the private cloud capacity for a decision horizon

p: the price of the public cloud per time unit

q: the guaranteed service level

k_a: the default downtime loss/penalty cost

t: the number of time units in a decision horizon

This section presents a base model for the cloud evaluation and investment decision problem. The following several assumptions used in this paper are from Henneberger [37]. It is assumed that private cloud resources are purchased or contracted at the beginning of the decision horizon, that the public cloud provider has a sufficient capacity to satisfy the peak demand of the company, and that the probability distribution for computing demand can be estimated based on real demand data and each demand can be split into the private and public clouds if cloud bursting is necessary. The unit price of the public cloud remains the same over the planning horizon.

Figure 1 shows a simulated demand distribution based on an exponential probability distribution. It is noted that a number of previous studies on models of computing resources assume the exponential distribution or the Weibull distribution of computing requests [11, 15, 16]. It is possible to use any probability distribution with real data, but no closed form solution may be obtainable and a simulation approach may be more appropriate over an analytical approach. x represents an actual computing demand and c represents the private cloud capacity purchased. When the actual computing demand x is below the purchased capacity c, the private cloud will be used. In this situation, $c\text{-}x$ is an excess private cloud capacity. When x is the same as c, the private cloud is fully utilized without the need for the public cloud. When the computing demand x exceeds c, then $x\text{-}c$ units of the public cloud are purchased from a public cloud provider via cloud bursting. In this model, the decision variable is the units of private cloud capacity to be purchased, c. Among the model variables, critical model variables include a probability distribution function for the computing demand, the price of the public cloud, and the purchase cost per unit of the private cloud.

The base model

The objective of the base model is to decide the optimal private cloud capacity to minimize the total cloud costs.

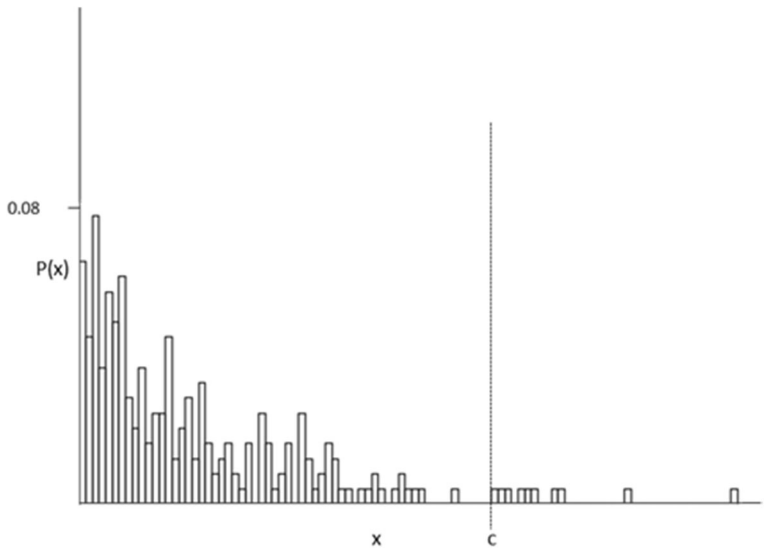

Fig. 1 An Example of a Demand Distribution based on an Exponential Probability Distribution

The base model of the hybrid cloud is given as the total cost minimization function (1).

$$\text{Min } TC = k \cdot c + \left(q \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot p + (1-q) \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot k_a \right) \cdot t \quad (1)$$

where $k \cdot c$ is the total private cloud costs at the beginning of the investment horizon, $q \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot p$ is the costs of the public cloud per time unit, and $(1-q) \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot k_a$ is the default downtime penalty cost related to the out-of-service situation from the public cloud provider. The cost structure follows Henneberger [37]. Applying integration techniques, Eq. (1) is transformed into Eq. (2).

$$TC = k \cdot c + (qpt + (1-q)k_a t)\left(\frac{1}{\lambda} e^{-\lambda c}\right) \quad (2)$$

Differentiating Eq. (2) in terms of c leads to:

$$\frac{dTC}{dc} = k - e^{-\lambda c}(qpt + (1-q)k_a t) \quad (3)$$

To get the closed form optimal value of c, Eq. (3) is set to zero. Then, the optimal private cloud capacity required is:

$$c^* = \frac{\ln\left(\frac{k}{(qpt + (1-q)k_a t)}\right)}{-\lambda} \quad (4)$$

Equation (5) is the second derivative of Eq. (2). As the second derivative is always greater than zero for the value of c, the objective function (2) is convex with a minimum at c^*.

$$\frac{d^2 TC}{dc^2} = \lambda e^{-\lambda c}(qpt + (1-q)k_a t) \quad (5)$$

The utilization rate of the private cloud c is given by:

$$u = \frac{-\frac{1}{\lambda} e^{-\lambda c} + \frac{1}{\lambda}}{c} \quad (6)$$

An illustration of the base model operation

As an illustration of the model operation, assume the following: $\lambda = 0.001$; $k = \$10{,}000$; $p = \$1.9$; $q = 0.9945$; $k_a = \$100$; $t = 10{,}000$. In this scenario, the unit price of the private cloud capacity is set to about 53% of the equivalent usage of the public cloud (i.e., \$10,000/(\$1.9*10,000)). This pricing assumption reflects the current cloud market. As of January 2018, Microsoft Azure provides cost savings of 40% to 68% for an advanced purchase of a virtual machine for one or 3 years compared to the hourly-based pay-as-you-go services (see https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/). 451 Research [39] also reported significant cost savings with the private cloud. The expected demand is 1000 capacity units (e.g., virtual machines or physical machines) with the exponential probability ($\lambda = 0.001$). Then, the optimal private cloud capacity, c^*, is 891.8 units, and the total computing cost, TC, is \$18,918,135. The optimal utilization rate of the private cloud is 66.17%.

Analysis of model behaviors

This section analyzes the relationship between the capacity decision and model variables. The experiment considers a set of possible pricing scenarios and determines the optimal cloud capacity to minimize the total costs.

Figures 2 and 3 show that as the price of the private cloud increases, the optimal private cloud capacity decreases rapidly from over 1400 to less than 600, but the optimal utilization rate of the private cloud increases slowly from 0.54 to 0.77. It is conventional wisdom that a higher utilization rate is desirable. A high utilization rate is not necessarily desirable since it is possible to achieve the high utilization rate with an under-capacity investment in the private cloud. It is interesting to note that the slight improvement of the utilization rate comes with a much greater reduction rate of the private cloud capacity.

Figure 4 shows that as the price of the public cloud increases, the use of the public cloud decreases, but, the use of the private cloud increases to meet the computing demand. Note that the total usage of both private and public cloud is 1000 units. When the price of the public cloud is $1.46, the usage of the public cloud is approximately equal to the usage of the private cloud. It is interesting to note that the usage lines of the public and the private cloud have the same distance from the straight line of the 500 units.

An evaluation and investment in interoperability for cloud bursting

The hybrid cloud requires interoperability of the private and the public cloud to support cloud bursting. Since cloud providers often offer their own proprietary applications, interfaces, and infrastructures, users have difficulty in migrating to cloud providers [40]. Two major approaches were proposed to improve cloud interoperability: provider-centric and user-centric approaches [41]. The provider-centric approaches are driven by a service provider who offers specific services and is willing to develop standards and technologies for customers to achieve a specified level of interoperability between the private and public cloud. On the other hand, the user-centric approaches are driven by the cloud users who either rely on their own in-house IT personnel or a third party (e.g., a cloud broker) to achieve the desired level of interoperability. For example, cloud users may develop a separate layer to handle heterogeneity in cloud environments [42]. In the hybrid computing environment, it is also challenging to decide which applications should be transitioned into the public cloud in time of cloud bursting. Lowering the bursting time and automating the movement process at the proper time require interoperable cloud systems. However, existing commercial tools rely heavily on the system administrator's knowledge to answer key questions such as when a cloud burst is needed and which applications must be moved to the cloud [31].

One benefit of interoperable cloud systems is that it reduces the cost of cloud bursting for users. The interoperable cloud systems may be achieved by using standardized data formats, open source cloud systems, application program interface (API), and specialized middleware. All these require time and investment by the users and/or cloud providers and there may exist a diminishing return on the investments. For example, cloud bursting requires that the company's software is configured to run multiple instances simultaneously and they may need to retrofit existing applications to accommodate multiple instances [43]. Therefore, it is important to evaluate the value of the interoperability-enhancing technologies to support cloud bursting.

This section focuses on the user-centric approaches where a corporate cloud customer decides to invest in interoperability enhancement such as the development of a layer to handle cloud bursting. Mathematical procedures are developed to evaluate investment decisions for the hybrid cloud capacity and the interoperability enhancement simultaneously. The base model is extended to include the investment decision variable for the interoperability enhancement. The

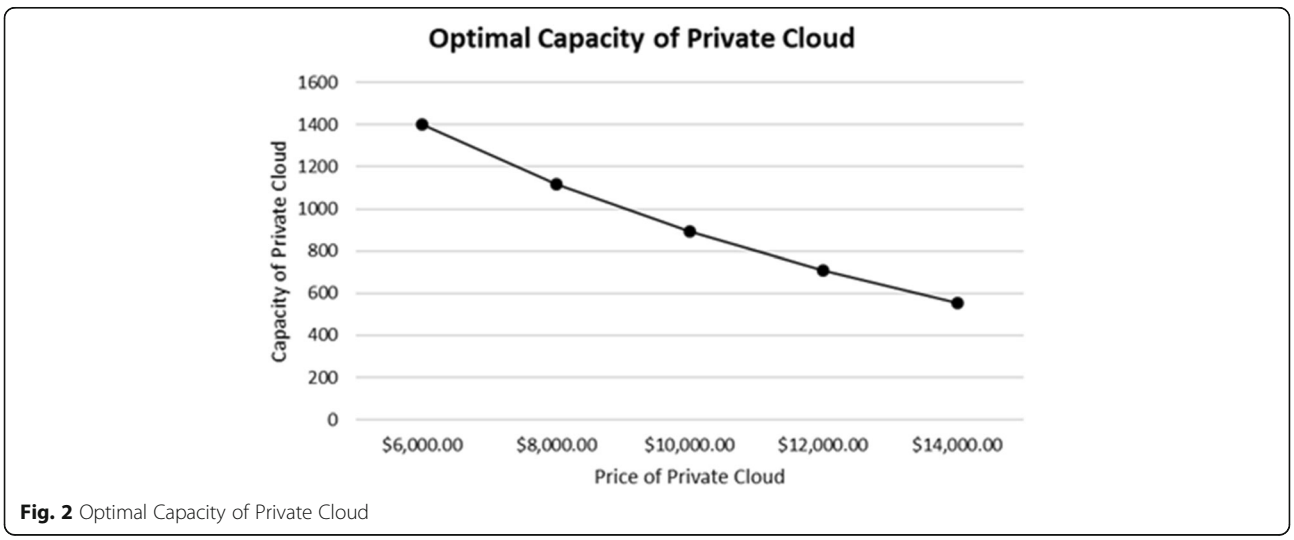

Fig. 2 Optimal Capacity of Private Cloud

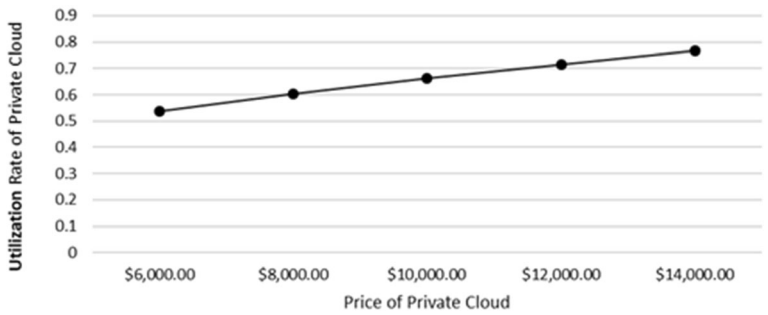

Fig. 3 Optimal Utilization Rate of Private Cloud

simultaneous cloud-interoperability decisions refer to the investment decisions for which both the variable for the cloud capacity and the variable for the interoperability enhancement are solved at the same time in the hope of reducing the total cloud costs synergistically. The cloud-interoperability evaluation and investment model is given as the total cost minimization function (7).

$$\text{Min } TC = k \cdot c + \left(q \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot p + q \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot z + (1-q) \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot k_a\right) \cdot t + G \quad (7)$$

Where z is a per-unit interoperability cost, $q \cdot \int_c^{\infty} \lambda e^{-\lambda x}(x-c)dx \cdot z$ is the total interoperability cost of the public cloud per time unit when cloud bursting occurs, and G is the investment to enhance the interoperability. Note that all other terms are the same as those in the base model. The total cost minimization function (7) is transformed to Eq. (8).

$$TC = k \cdot c + (qpt + qzt + (1-q)k_a t)\left(\frac{1}{\lambda}e^{-\lambda c}\right) + G \quad (8)$$

Now, the per-unit interoperability cost, z, is defined as an exponential function of the investment, G, as follows.

$$z = U + (W-U)e^{-\beta G}, \quad G \geq 0, \text{ and } 0 \leq U \leq z \leq W \quad (9)$$

where W is the highest interoperability cost per time unit of the public cloud incurred when there is no investment in the interoperability enhancement and U is the lowest interoperability cost achievable with the investment of G.

Differentiating Eq. (8) in terms of c and G, respectively leads to:

$$\frac{dTC}{dc} = k - e^{-\lambda c}(qpt + qzt + (1-q)k_a t) \quad (10)$$

$$\frac{dTC}{dG} = \left(qz't\right)\left(\frac{1}{\lambda}e^{-\lambda c}\right) + 1 \quad (11)$$

To simplify the mathematical procedures,
set $e^{-\lambda c} = g$
Equation (11) leads to:

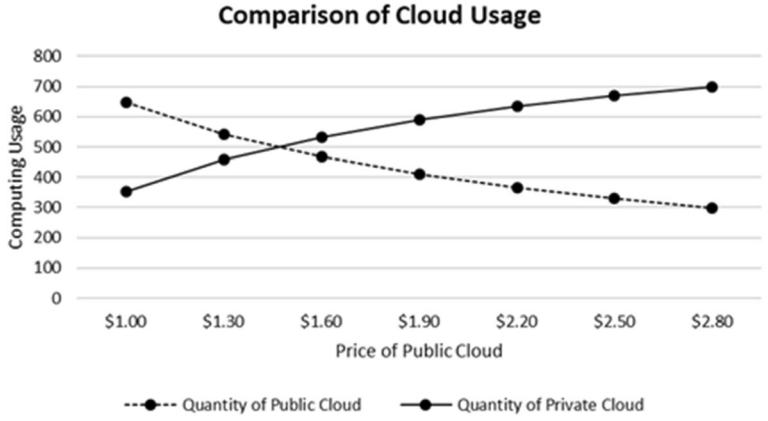

Fig. 4 Comparison of Cloud Usage

$$\frac{dz}{dG} = \frac{-\lambda}{qtg} \tag{12}$$

The first derivative of Eq. (9) is taken with regard to G. The result is given by:

$$\frac{dz}{dG} = -\beta(z-U) < 0 \tag{13}$$

Setting two Eqs. (12) and (13) equal leads to:

$$z = \frac{\lambda + U\beta qtg}{\beta qtg} \tag{14}$$

If the optimal private cloud capacity $c^*$is given, the optimal z^* is:

$$z^* = \frac{\lambda + U\beta qte^{-\lambda c^*}}{\beta qte^{-\lambda c^*}} \tag{15}$$

To find the simultaneous solution for both the private cloud capacity and the interoperability enhancement, set Eq. (10) to zero, and plug Eq. (14) in the equation to get:

$$g = \frac{(\beta k - \lambda)}{(\beta qpt + U\beta qt + \beta k_a t - \beta q k_a t)} \tag{16}$$

Given the optimal g^* (i.e., $e^{-\lambda c}$), z^*, c^*, and G^* are obtained as follows:

$$z^* = \frac{\lambda + U\beta qtg^*}{\beta qtg^*} \tag{17}$$

$$c^* = \frac{-\ln g^*}{\lambda} \tag{18}$$

$$G^* = \frac{-\left(\ln \frac{(z^*-U)}{(W-U)}\right)}{\beta} \tag{19}$$

Table 1 shows the improvement made by the simultaneous cloud-interoperability investment decision over the sequential cloud-interoperability investment decision and no investment decision. The sequential cloud-interoperability investment decision refers to the evaluation process in which only one decision variable is solved at a time. First, using Eq. (4), the optimal private cloud capacity, c^*, is determined without considering the optimal interoperability investment, G^*. Second, based on the optimal private cloud capacity, using Eq. (15), the investment in the optimal interoperability enhancement is determined. For example, assume the same base model's parameters. Then, the optimal private cloud capacity $c^* \approx 435$. Then, $g = e^{-\lambda c} = e^{(-0.001 * 435)} = 0.64746$. The optimal interoperability cost, z^*, is \$0.4883. The optimal investment in the interoperability enhancement, G^*, is \$1,807,338. Table 1 shows that the sequential cloud-interoperability investment decision results in a lower interoperability cost per computing unit than the simultaneous cloud-interoperability investment decision, but it brings about overinvestment in the interoperability enhancement and consequently larger total cloud costs, as it does not take into consideration the synergy effect. On the other hand, the simultaneous cloud-interoperability investment decision leads to an optimal investment, considering the synergy effect between the interoperability cost and the cloud capacity cost. Surprisingly, the sequential interoperability investment generates a higher total cost than no interoperability investment.

Sensitivity analysis of cloud-interoperability investment decision

This section further evaluates the performance of the sequential and simultaneous investment approaches discussed above and conducts sensitivity analyses to understand model behaviors of the investment decisions by changing parameter values. The base parameter values assumed are presented below.

Base parameters

k: a cost of \$10,000 per unit private cloud capacity
p: the unit price of \$1.0 for public cloud service per time unit
z: the interoperability cost of \$0.9 without investment
q: the guaranteed service level of 99.45%
k_a: a default downtime loss/penalty cost of \$100 per time unit
t: 10,000 time units in a decision horizon
U = \$0.1
W = \$0.9
λ = 0.001
β = 0.0000004

Table 1 Comparison of investment approaches

	No Interoperability Investment (A)	Sequential Interoperability Investment (B)	Simultaneous Interoperability Investment (C)	Improvement (A-C)	Improvement (B-C)
TC^*	\$18,918,136	\$19,298,239	\$18,779,969	\$138,167	\$518,270
z^*	\$0.9000	\$0.4883	\$0.6510	\$0.2490	-\$0.1628
c^*	892	435	785	107	−350
G^*	\$0	\$1,807,338	\$932,129	-\$932,129	\$875,209

Figures 5 shows that as the computing demand increases, the optimal interoperability cost, z^*, decreases with more investment in the interoperability enhancement, G. The decline of the per-unit interoperability cost is more rapid when the demand is smaller and the decline slows down as the demand increases.

Figure 6 shows that as the demand level increases, the difference gets wider between the total cost of the simultaneous cloud-interoperability and that of the no investment decision. The simultaneous cloud-interoperability dominates no investment at all demand levels. Figure 7 shows that the total cost of the sequential cloud-interoperability investment is higher than that of the no investment decision when the demand level is small, but becomes lower than that of the no investment decision as the demand gets larger.

The analysis indicates that an organization with a large cloud demand level would achieve a greater cost advantage from the interoperability enhancement than an organization with a small demand, since a marginal increase of the investment in the interoperability enhancement generates large savings for the total cloud cost when the demand is high. On the other hand, many small and medium-sized companies may find the use of a third-party cloud broker to be more cost effective than an in-house development of interoperability technologies. Small and medium-sized companies may also reap benefits from cloud providers which offer standardized interoperable cloud technologies.

Return on investment of an interoperability project

As in many other IT projects, one of the barriers to the investment in cloud technologies is difficulty in measuring the potential return on investment (ROI). ROI is one of the most crucial criteria for companies to consider when investing in a new technology. ROI analyses have been used in various technology evaluations [44]. For example, Lee [45] develops cost/benefit models to measure the impact of RFID-integrated quality improvement programs and to predict the ROI. Using these models, the decision makers decide whether and how much to invest in quality improvement projects. An ROI model was developed for an economic evaluation of cloud computing investment [27].

While a formula for calculating ROI is relatively straightforward, the ROI method is more suitable when both the benefits and the costs of an investment are easily traceable and tangible. Since an investment in cloud technologies is a capital expenditure, the investment is likely to be under scrutiny of senior management for the budget approval. The investment evaluation and decision model in the previous sections identifies the optimal investment point where the marginal cost saving equals the marginal investment. However, it is possible that the optimal solution does not necessarily meet the ROI threshold rate imposed by the organization. The ROI threshold rate is the minimum ROI level for which a company will make investments. The ROI threshold rate is usually tied to the cost of capital and is used to screen out low productivity projects under financial constraints. This section derives a formula to identify the target ROI of the interoperability enhancement project.

To illustrate the calculation of the ROI, suppose the total cloud cost without investment in interoperability is $100 m. Assume that with the investment of $10 million, the total cloud cost including the investment is $99 m. In this case the ROI is 10% (i.e., ($100 m-$99 m)/$10 m). If the ROI threshold rate of the company is 20%, the project would be rejected even though the investment of $10 million reduces the total cloud cost. The basic ROI formula is given as:

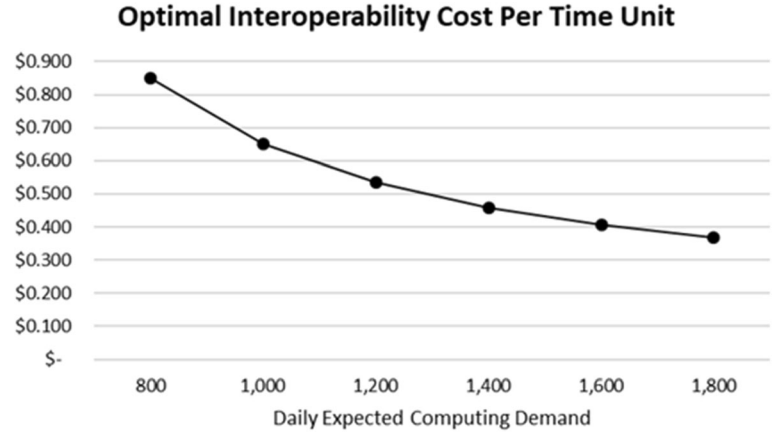

Fig. 5 Change of Computing Demand and and Optimal Interoperability Cost

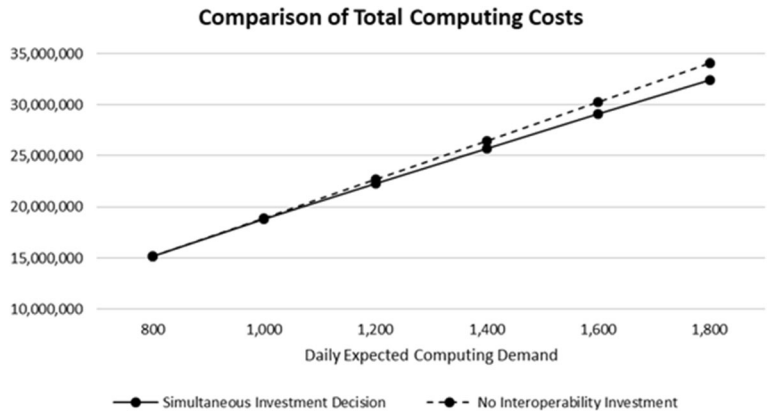

Fig. 6 Change of Computing Demand and Total Costs by Simultaneous Investment and No Investment

$$ROI = \left(\frac{Total\ Cost\ without\ investment - Total\ cost\ with\ investment}{Invesment}\right) \cdot 100 \tag{20}$$

Since the optimal investment, G^*, is the point where the marginal cost saving equals the marginal investment, it is expected that the decrease of the investment amount increases the ROI. Equation (21) is used to measure the ROI.

$$\frac{I - \left(k \cdot c + (qpt + qzt + (1-q)k_a t)\left(\frac{1}{\lambda}e^{-\lambda c}\right) + G\right)}{G} = r \tag{21}$$

where I is the total cloud cost without investment in interoperability and r is the ROI target threshold.

To find the target ROI using the Newton-Raphson method in one variable, the following is formulated:

$$f(G) = I - \left(k \cdot c + (qpt + qzt + (1-q)k_a t)\left(\frac{1}{\lambda}e^{-\lambda c}\right) + G\right) - rG \tag{22}$$

Then, the first derivative of $f(G)$ is:

$$f'(G) = -\left(qz't\right)\left(\frac{1}{\lambda}e^{-\lambda c}\right) - 1 - r \tag{23}$$

where $z' = -\beta(W - U)e^{-\beta G}$.

To begin the search for the target ROI, r, set G_0 at G^* found in Eq. (19).

$$G_1 = G_0 - \frac{f(G_0)}{f'(G_0)} \tag{24}$$

The process is repeated as:

$$G_{n+1} = G_n - \frac{f(G_n)}{f'(G_n)} \tag{25}$$

until a sufficiently accurate value of G is reached.

As an illustration of the above Newton-Raphson method, assume that the target ROI is 20%. Using the

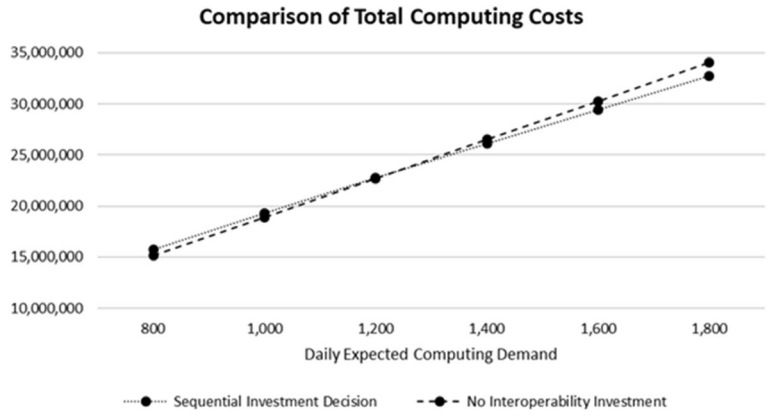

Fig. 7 Change of Computing Demand and Total Costs by Sequenctial Investment and No Investment

base parameters given in "Sensitivity analysis of cloud-interoperability investment decision" section, G_0 is set at $932,129. G_1 is $690,836 and ROI is 14.83%. G_2 is $620,350 and ROI is 18.72%. G_3 is $612,780 and ROI is 19.87%. G_4 is $612,690 and ROI is 19.9985%. After only four iterations of Eqs. (24) and (25), a sufficiently accurate value of G is obtained to achieve the target ROI of 20%. Compared to a time-consuming linear search for the G for the target ROI, the Newton-Raphson method is shown to achieve tremendous efficiency in the cloud investment decision and can be a useful performance management technique.

A sensitivity analysis is conducted to investigate the relationship between the ROI and G. Note that to minimize the total cloud cost, the optimal investment should be $932,129 and the ROI is 14.82%. Figure 8 shows that the decrease of the investment from $932,129 generates higher ROIs, but the increase of the investment from $932,129 decreases the ROIs. If the ROI threshold rate of the company is higher than 14.82%, they need to decrease the investment even though the total cloud cost is the lowest at the investment of $932,129. With the ROI, managers would be able to develop a strong justification for the investment.

Diminishing return applies to the ROI of the interoperability investment. A corporate cloud customer can find that for their given computing demand distribution, there is an optimal level of interoperability investment that minimizes the total cloud cost. Up to the optimal level of investment, any additional investment will increase the return. However, over that optimal level of investment, any additional investment will continue to diminish the return. The corporate cloud customer may have an ROI threshold rate which is equal to the company's minimum acceptable ROI rate. If the ROI at the optimal total cloud cost is greater than or equal to the ROI threshold rate, then the interoperability investment level needed for the optimal total cloud cost is justified. However, if the ROI at the optimal total cloud cost is less than the ROI threshold rate, the company needs to lower the interoperability investment to the point where the ROI equals their ROI threshold rate.

Conclusion

According to Forrester [46], in 2018, cloud computing became a must-have technology for every enterprise. Nearly 60% of North American enterprises are using some type of public cloud platform. Furthermore, private clouds are also growing fast, as companies not only move workloads to the public cloud but also develop on-premises private cloud in their own data centers. Therefore, this paper predicts that the corporate adoption of hybrid cloud computing for corporate cloud capacity management is an irreversible trend. The demand for the hybrid cloud will continue to grow in the future as big data, smart phones, and the Internet of Things (IoT) technologies require highly scalable infrastructure to meet the growing but fluctuating computing demand.

Advances in cloud computing hold great promise for lowering computing costs and speeding up new business developments. However, despite the popularity of cloud computing, no significant investment evaluation models are available. The existing body of studies are mostly descriptive in nature. The optimal cloud capacity investment has been elusive and therefore investments have been based on gut feelings rather than solid decision models. Hence, this paper presented a capacity evaluation and investment decision model for the hybrid cloud. A closed form solution was derived for the optimal capacity decision. This model provides a solid foundation to evaluate investment decisions for various cloud technologies, moving beyond the existing descriptive valuation studies to normative studies, the outcomes of which should be able to guide cloud professionals to plan on how much they need to invest in different cloud technologies and how they can enhance the investment benefits from a corporate customer's perspective.

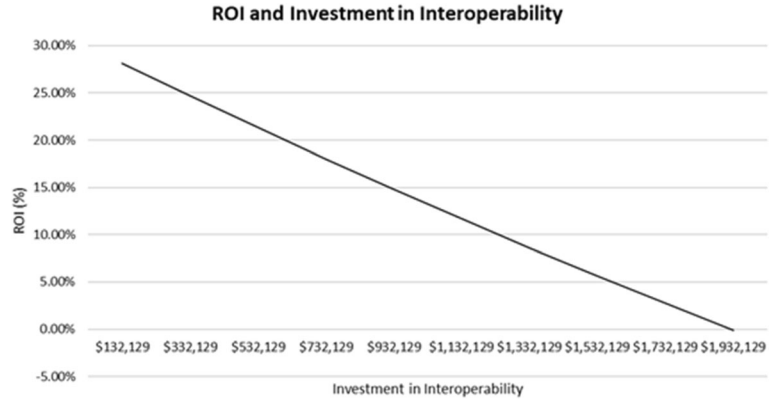

Fig. 8 Change of the ROI over the Investment in Interoperability

While there is widespread cloud adoption and technological breakthrough, the interoperability issue remains a barrier. To address the interoperability issue in the hybrid cloud, this paper also developed the cloud-interoperability evaluation and investment model by extending the basic model. This model derives the optimal investment levels for both the cloud capacity and the interoperability enhancement simultaneously. The relationship between the investment in the interoperability enhancement and ROI was also investigated.

This study is the first effort in developing an analytical cloud capacity evaluation and investment decision model for the hybrid cloud, and analyzing the impact of the interoperability investment on the cost savings and ROI. This model can be easily adapted to a variety of investment situations in both for-profit organizations and non-profit organizations such as hospitals, governments, and libraries that have similar computing demand and cost structures.

Like many studies, this study has several limitations. First, it is noted that the successful use of these models would require accurate estimation of the model parameters and the use of various modeling techniques. Future research may explore various estimation techniques of the model parameters to develop more realistic models. Second, future studies need to investigate additional variables for the model. For example, different cloud providers employ different schemes and models for pricing [47] and the diversity in the pricing models makes price comparison difficult [48]. Incorporation of this variable price may be a worthwhile future research. The interoperability costs may include not only tech architecture or wrappers, but also data transport costs. The future model may include the data transport costs in the cloud capacity decision. Lastly, despite an increased focus on cloud cost management, only a minority of companies have implemented policies to minimize wasted cloud resources such as shutting down unused workloads or rightsizing instances [5]. Implementation of well-defined cloud governance and adaptive cloud scheduling techniques may help companies minimize the wasted cloud resources and avoid overinvestment in cloud capacity. Researchers are encouraged to explore the above-mentioned limitations of this study in their future research.

Abbreviations
RFID: Radio Frequency Identification; ROI: Return on investment

Acknowledgements
Not applicable.

Authors' contributions
In Lee is the sole author of this study. The author read and approved the final manuscript.

Authors' information
Not applicable.

Funding
Not applicable.

Availability of data and materials
Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests
The author declares that he has no competing interests.

Received: 29 March 2019 Accepted: 15 October 2019
Published online: 01 November 2019

References
1. Varghese B, Buyya R (2018) Next generation cloud computing: new trends and research directions. Future Gener Comput Syst 79(Part 3):849–861
2. Birje MN, Challagidad PS, Goudar RH, Tapale MT (2017) Cloud computing review: concepts, technology, challenges and security. Int J Cloud Comput 6(1):32–57
3. Hentschel R, Leyh C, Petznick A (2018) Current cloud challenges in Germany: the perspective of cloud service providers. J Cloud Comput 7:5. https://doi.org/10.1186/s13677-018-0107-6
4. Opara-Martins J, Sahandi R, Tian F (2016) Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. J Cloud Comput Adv Syst Appl 5:4. https://doi.org/10.1186/s13677-016-0054-z
5. RightScale (2019). State of the cloud report. Available from: https://www.rightscale.com/lp/state-of-the-cloud?campaign=RS-HP-SOTC2019
6. Lee I (2017) Determining an optimal mix of hybrid cloud computing for enterprises. UCC '17 companion. In: Proceedings of the10th international conference on utility and cloud computing, pp 53–58
7. Azumah KK, Sørensen LT, Tadayoni R (2018) Hybrid cloud service selection strategies: a qualitative meta-analysis. In: 2018 IEEE 7th international conference on adaptive science & technology (ICAST), Accra, Ghana
8. Litoiu M, Wigglesworth J, Mateescu R (2016) The 8th CASCON workshop on cloud computing. In: Proceeding CASCON '16 proceedings of the 26th annual international conference on computer science and software engineering, pp 300–303
9. Computerweekly (2018) Available from: https://www.computerweekly.com/news/252452059/Uber-backs-hybrid-cloud-as-route-to-business-and-geographical-expansion. Accessed 21 January 2019.
10. Alashoor T (2014) Cloud computing: a review of security issues and solutions. Int J Cloud Comput 3(3):228–244
11. Khazaei H, Misic J, Misic VB (2011) Modelling of cloud computing centers using M/G/m queues. In: 2011 31st international conference on distributed computing systems workshops, Minneapolis, MN, USA, pp 87–92
12. Moreno I, Garraghan P, Townend P, Xu J (2013) An approach for characterizing workloads in google cloud to derive realistic resource utilization models. In: Proceedings of 7th international symposium on service oriented system engineering (SOSE), Redwood City, CA, pp 49–60
13. Magalhães D, Calheiros RN, Buyya R, Gomes DG (2015) Workload modeling for resource usage analysis and simulation in cloud computing. Comput Electr Eng 47:69–81
14. Rochman Y, Levy H, Brosh E (2017) Dynamic placement of resources in cloud computing and network applications. Perform Eval 115:1–37
15. Gupta V, Harchol-Balter M, Sigman K, Whitt W (2007) Analysis of join-the-shortest-queue routing for web server farms. Perform Eval 64(9–12):1062–1081
16. Yang B, Tan F, Dai Y, Guo S (2009) Performance evaluation of cloud service considering fault recovery. In: First Int'l conference on cloud computing CloudCom, Beijing, China, pp 571–576
17. Xiong K, Perros H (2009) Service performance and analysis in cloud computing. In: IEEE 2009 world conference on services, Los Angeles, CA, pp 693–700
18. Patch B, Taimre T (2018) Transient provisioning and performance evaluation for cloud computing platforms: a capacity value approach. Perform Eval 118:48–62
19. Wolski R, Brevik J (2014) Using parametric models to represent private cloud workloads. IEEE Trans Serv Comput 7(4):714–725
20. Grid Workloads Archive (2017) Available from: http://gwa.ewi.tudelft.nl/. Accessed 21 January 2019.
21. Araujo J, Maciel P, Andrade E, Callou G, Alves V, Cunha P (2018) Decision making in cloud environments: an approach based on multiple-criteria decision analysis and stochastic models. J Cloud Comput Adv Syst Appl 7:7. https://doi.org/10.1186/s13677-018-0106-7
22. López-Pires F, Barán B (2017) Cloud computing resource allocation taxonomies. Int J Cloud Computing 6(3):238–264
23. Balaji M, Kumar A, Rao SVRK (2018) Predictive cloud resource management framework for enterprise workloads. J King Saud Univ Comput Inf Sci 30(3):404–415

24. Wang CF, Hung WY, Yang CS (2014) A prediction based energy conserving resources allocation scheme for cloud computing. In: 2014 IEEE international conference on granular computing (GrC), Noboribetsu, Japan, pp 320–324
25. Han Y, Chan J, Leckie C (2013) Analysing virtual machine usage in cloud computing. In: 2013 IEEE ninth world congress on services, Santa Clara, CA, USA, pp 370–377
26. Deng D, Lu Z, Fang W, Wu J (2013) CloudStreamMedia: a cloud assistant global video on demand leasing scheme. In: 2013 IEEE international conference on services computing, Santa Clara, CA, USA, pp 486–493
27. Misra SC, Monda A (2011) Identification of a company's suitability for the adoption of cloud computing and modelling its corresponding return on investment. Math Comput Model 5(3/4):504–521
28. de Assunção MD, di Costanzo A, Buyya R (2010) A cost-benefit analysis of using cloud computing to extend the capacity of clusters. Clust Comput 13(3):335–347
29. Gmach D, Rolia J, Cherkasova L, Kemper A (2007) Capacity management and demand prediction for next generation data centers. In: IN: IEEE international conference on web services, Salt Lake City, UT, USA, pp 43–50
30. Laatikainen G, Mazhelis O, Tyrvainen P (2016) Cost benefits of flexible hybrid cloud storage: mitigating volume variation with shorter acquisition cycle. J Syst Softw 122:180–201
31. Guo T, Sharma U, Shenoy P, Wood T, Sahu S (2014) Cost-aware cloud bursting for enterprise applications. ACM Trans Internet Technol (TOIT) 13(3):10 22 pages
32. Clemente-Castelló FJ, Mayo R, Fernández JC (2017) Cost model and analysis of iterative MapReduce applications for hybrid cloud bursting. In: CCGrid '17 Proceedings of the 17th IEEE/ACM international symposium on cluster, cloud and grid computing, Madrid, Spain, pp 858–864
33. Weinman J (2016) Hybrid cloud economics. IEEE Cloud Comput 3(1):18–22
34. Toosi AN, Sinnott RO, Buyya R (2018) Resource provisioning for data-intensive applications with deadline constraints on hybrid clouds using Aneka. Future Gener Comput Syst 79(Part 2):765–775
35. Petcu D (2011) Portability and interoperability between clouds: challenges and case study. In: Abramowicz W, Llorente IM, Surridge M, Zisman A, Vayssière J (eds) Towards a service-based internet. ServiceWave 2011. Lecture notes in Computer Science, vol 6994. Springer, Berlin, pp 62–74
36. Edmonds A, Metsch T, Papaspyrou A, Richardson A (2012) Toward an open cloud standard. IEEE Internet Comput 16(4):15–25
37. Henneberger M (2016) Covering peak demand by using cloud services - an economic analysis. J Decis Syst 25(2):118–135
38. Fan TJ, Chang XY, Gu CH, Yi JJ, Deng S (2014) Benefits of RFID technology for reducing inventory shrinkage. Int J Prod Econ 147(Part C):659–665
39. 451 Research (2017) Can private cloud be cheaper than public cloud? Available from: https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/vrealize-suite/vmware-paper1-can-private-cloud-be-cheaper-than-public-cloud.pdf
40. Mansour I, Sahandi R, Cooper K, Warman A (2016) Interoperability in the heterogeneous cloud environment: a survey of recent user-centric approaches. In: ICC '16 proceedings of the international conference on internet of things and cloud computing, Article No. 62, Cambridge: ACM, New York.
41. Toosi AN, Calheiros RN, Buyya R (2014) Interconnected cloud computing environments: challenges, taxonomy, and survey. ACM Comput Surv 47(1):7 47 pages
42. Zhang Z, Wu C, Cheung DWL (2013) A survey on cloud interoperability: taxonomies, standards, and practice. ACM SIGMETRICS Perform Eval Rev 40(4):13–22
43. Morpheus (2016) If cloud bursting is so great, why aren't more companies doing it? Available from: https://www.morpheusdata.com/blog/2016-10-26-if-cloud-bursting-is-so-great-why-aren-t-more-companies-doing-it
44. Sarac A, Absi N, Dauzère-Pérès S (2010) A literature review on the impact of RFID technologies on supply chain management. Int J Prod Econ 128(1):77–95
45. Lee HH (2008) The investment model in preventive maintenance in multi-level production systems. Int J Prod Econ 112(2):816–828
46. Forrester (2018) Predictions 2019: Cloud computing comes of age as the foundation for enterprise digital transformation. Available from: https://go.forrester.com/blogs/predictions-2019-cloud-computing/
47. Al-Roomi M, Al-Ebrahim S, Buqrais S, Ahmad I (2013) Cloud computing pricing models: a survey. Int J Grid Distr Comput 6(5):93–106
48. Lehmann S, Buxmann P (2009) Pricing strategies of software vendors. Bus Inf Syst Eng 1(6):452–462

Publisher's Note