A preprint version of the manuscript
to appear in Information Systems Frontiers
The final publication is available at www.springerlink.com

Economic Aspects of Hybrid Cloud Infrastructure: User Organization Perspective

Oleksiy Mazhelis · Pasi Tyrväinen

Abstract Adoption of cloud infrastructure promises enterprises numerous benefits, such as faster time-to-market and improved scalability enabled by on-demand provisioning of pooled and shared computing resources. In particular, hybrid clouds, by combining the private in-house capacity with the on-demand capacity of public clouds, promise to achieve both increased utilization rate of the in-house infrastructure and limited use of the more expensive public cloud, thereby lowering the total costs for a cloud user organization. In this paper, an analytical model of hybrid cloud costs is introduced, wherein the costs of computing and data communication are taken into account. Using this model, a cost-efficient division of the computing capacity between the private and the public portion of a hybrid cloud can be identified. By analyzing the model, it can be shown that, given fixed prices for private and public capacity, a hybrid cloud incurs the minimum costs. Furthermore, it is shown that, as the volume of data transferred to/from the public cloud increases, a greater portion of the capacity should be allocated to the private cloud. Finally, the paper illustrates analytically that, when the unit price of capacity declines with the volume of acquired capacity, a hybrid cloud may become more expensive than a private or a public cloud.

Keywords hybrid cloud · cost model · cost optimization · price elasticity · steepness of quantity discount

1 Introduction

Cloud computing represents a state-of-the-art "computing as a service" paradigm, where configurable computing resources are pooled and shared among multiple

Oleksiy Mazhelis · Pasi Tyrväinen
Department of Computer Science and Information Systems
Agora, P.O.Box 35, FI-40014, University of Jyväskylä, Finland
E-mail: {oleksiy.ju.mazhelis, pasi.t.tyrvainen}@jyu.fi

users and efficiently provisioned to them, on-demand, through a broadband network access (Mell and Grance, 2010). The deployment of cloud infrastructure promises enterprises numerous benefits, such as faster time to market and improved scalability (Youseff et al, 2008), as well as cost benefits in terms of lower start-up and/or operations costs (Weinman, 2011; Lee, 2010). Due to these benefits, the adoption of cloud infrastructure services has intensified in recent years: according to Gartner, the market for cloud services exceeded $46 billion in 2009 and will reach $150 billion by 2013 (Pring et al, 2009).

According to Mell and Grance (Mell and Grance, 2010), cloud infrastructure may be deployed in a form of a private, community, public, or hybrid cloud. A private cloud is operated by a single organization, whereas a community cloud is shared and jointly operated by several organizations. These two deployment options are justified, either when the computing needs are large, or when the demand is relatively flat. In contrast, a public cloud is operated by an independent cloud service provider; this mode is attractive, e.g. to small user organizations, enabling them to avoid large up-front IT investments. A hybrid cloud is a combination of a public cloud and a private cloud, and is aimed at providing an efficient distribution of the load among the clouds.

In the case of a hybrid cloud, complementing the local infrastructure with computing capacity from a public cloud enables organizations to increase the utilization of their IT infrastructure and thereby reduce their IT costs. As argued by Weinman (Weinman, 2011), a hybrid cloud is more cost-efficient than a private cloud, since the high premium charged by the public cloud provider is compensated by the relatively short duration of load peaks when the public cloud is utilized. Furthermore, when a load is uniformly distributed between zero and maximum during an observed time period, the cost-optimum portion of the public cloud load is the inverse of the premium charged by the cloud service provider (Weinman, 2011). The cost-optimal load distribution in (Weinman, 2011) assumes that only the computing capacity is charged for by the cloud service provider, and that no other costs affect the analysis. This is not the case, however, in many data-intensive applications, where a significant volume of data needs to be transferred to/from the cloud, thereby incurring data communication costs (Mazhelis and Tyrvinen, 2011).

The cost advantages of a hybrid solution are partially confirmed in (Risch and Altmann, 2008), where the conclusion made is that the usage of a computing grid infrastructure is economically advantageous when the demand for computing exhibits infrequent (in intervals of several months) peaks that can be covered with grid capacities. Different results have been obtained by (Strebel and Stage, 2010), who explicitly focus on the cost-efficient mix of internal and external computing resources in a hybrid cloud. In their approach, individual applications are assigned to either internal or external resources, using mixed-integer programming. Based on their simulation results, the authors have found that the off-loading peak de-

mand to the public cloud may not bring any cost-benefits to the clients, though the authors acknowledge the preliminary nature of the findings and suggest that there is a need for further research in this direction. The strength of the model is in the possibility to find a cost-optimal solution directing the assignment of applications to the resources. However, due to the nature of the optimization problem, as stated in (Strebel and Stage, 2010), the solution output delivered by the model is difficult to interpret, and hence its generalization to other environments is challenging, too.

The concurrent use of in-house and external capacity has been also a subject of extensive research outside of the information systems and computer science domains. In particular, the related phenomena of tapered integration (Porter, 1980), plural governance (Heide, 2003), and concurrent sourcing (Parmigiani, 2007) have been studied in organization and strategic management literature; see (Mols, 2010) for a comprehensive review. In these studies, the concurrent use of internal and external capacity has been considered from the viewpoint of different theories, including, among others, the transaction costs theory, the agency theory, the resource-based theory, and the theories of neo-classical economics, and numerous hypotheses explaining such concurrent use have been derived and empirically tested. In particular, in line with the principles of the neo-classical economics, it was found that in markets characterized by demand uncertainty, the risk of diseconomies of scale due to unutilized excess capacity may be mitigated by scaling down internal capacity and supplementing it during peak demand with externally acquired capacity (Heide, 2003; Puranam et al, 2006). However, the cost-efficient division between the concurrently used in-house and external capacity is considered in (Puranam et al, 2006) on a general level, and therefore it does not capture the specifics present in the concurrent use of the computing, storage, and communication capacities provided by the cloud infrastructure. Volume uncertainty is also one of the concepts considered in the transaction cost theory (Williamson, 1985), which predicts that firms facing volume uncertainty will likely rely on internal rather than external capacity. However, as discussed, e.g., in (Mols, 2010), the transaction cost economics, while focusing on the firms' choice between the use of internal and external capacity, does not explain the phenomenon of concurrent sourcing.

This paper aims at addressing the issue of efficient division of the load between the private and the public portion of a hybrid cloud. An analytical model of hybrid cloud costs, including the costs of computing and data communication, is introduced in the paper. In the analytical model, two phenomena that may affect the costs of using a hybrid cloud infrastructure are considered:

- *Variable demand for a particular resource capacity.* If there are peaks in the demand, in-house provisioning often leads to over-provisioning and under-utilized resources (Weinman, 2011).
- *Declining unit price of capacity, as the volume of acquired resources grows.* The more the resource capacity is concentrated in one place (in-house or a public

cloud), the cheaper the price of one unit of the resource due to the all-unit or incremental price discounting (Stole, 2003; Schotanus et al, 2009).

Using the model, the cost-optimal load division can be identified, as exemplified in the paper for the case of demand uniformly distributed between zero and maximum. It is shown analytically that, given an arbitrary demand distribution and fixed unit prices, a hybrid cloud provides the minimum costs; furthermore, the presence of data communication costs shifts the cost-optimal division towards the private cloud, i.e. the greater the data communication volume, the greater the portion of the demand that should be allocated to the private cloud. It is also analytically shown that when the price is subject to a quantity discount the hybrid cloud may become more expensive than a private and/or a public cloud.

Thus, this paper contributes to the previous work in the domain of the economics of cloud computing by introducing the cost model for a hybrid cloud infrastructure taking into account i) variable demand for computing capacity, ii) data communication overheads and iii) quantity discounts for the unit prices. The remainder of the paper is organized as follows. In the next section, a simplified architectural description of a hybrid cloud is provided, the relevant costs are defined and the main assumptions made are specified. The analytical model is introduced in Section 3, and its properties are analytically analyzed in Section 4. Numerical experiments illustrating the effect of data communication costs are described in Section 5. In section 6, the implications of the proposed model are discussed, and the directions for further work are outlined. Finally, conclusions to the paper are given in Section 7.

2 Hybrid Cloud

Throughout this paper, we will consider the case of a hybrid cloud, where a private and the public clouds are used in combination by an organization in order to provide service(s) to its customers. Let us assume that a portion of the organization's software can be deployed in a cloud, either private or public, while the other software subsystems, e.g. legacy subsystems, applications with strict performance requirements, or subsystems dealing with highly confidential data, have to be deployed in-house either using a traditional IT infrastructure or a private cloud. Thus, the overall software system architecture can be decomposed into three subsystems:

- The open subsystems provided by the public cloud;
- The open subsystems provided by the private cloud;
- The closed subsystems.

This decomposition is depicted in Fig. 1. The term *open subsystem* is employed to emphasize the fact that the subsystem deployment is not tied to the in-house infrastructure and can easily be changed from private to public cloud and back,

depending on the day-to-day management decisions. On the other hand, the closed subsystems are to be deployed in-house in the foreseeable future.

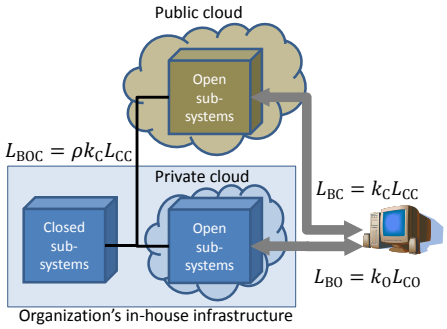

Fig. 1 System decomposition to its subsystems. The terms shown in the figure are introduced later in the text and denote the following: L_{bo}, L_{bc} and L_{boc} are the volumes of data transferred between the private cloud and the customers, between the public cloud and the customers and between the organization and the public cloud, respectively; L_{co} and L_{cc} denote the cumulative reserved private computing capacity and acquired public cloud computing capacity, respectively; k_o, k_c, and ρ are coefficients.

It is assumed that the same software is used in both the private and the public cloud subsystems. In case the software subsystems are heterogeneous, an ontology mapping can be employed to enable their interoperation and to make the system scalable (Jung, 2010).

Let us assume that the open subsystems are responsible for (a part of) information exchange with the customers, and instantiated, e.g., in a form of a web-portal, a content-distribution server, etc. Furthermore, let us assume that the interaction between the service side and the customer side requires a substantial volume of data to be transferred, as depicted in the figure by bold arrows.

The demand for the system's computing capacity is assumed to change in time. The demand up to a specific threshold value is supplied using the private cloud capacity, which is acquired beforehand and reserved for the purposes of service provisioning. Whenever the demand exceeds the threshold value, the private cloud capacity is no more sufficient to meet the demand, and the portion of the demand exceeding the threshold is supplied by the public cloud infrastructure, which is used without prior reservation and charged based on the actual usage.

An example of such system is an online image processing system[1] allowing the users to upload their images, edit them on-line, and then download the edited version. Private cloud subsystems are responsible for serving the users' edition requests coming at a regular rate; however, during the periods with heavy load,

[1] Such as pixIr (`http://pixlr.com/`) or Adobe Photoshop Express (`http://www.photoshop.com/tools`)

part of the requests are processed by public cloud subsystems. The responsibilities of the closed subsystems may include supporting functions, such as service activation, identity management, service level monitoring, charging and billing.

The proposed model is focused on identifying the distribution of open subsystems' computing capacity between the private and the public cloud, which would minimize the costs. The costs of closed subsystems are assumed to be independent of how the open subsystems are distributed, and therefore their costs are not taken into account when seeking the cost-optimal private-public cloud distribution. Thus, only the costs of open subsystems (private and public cloud) are analyzed. Furthermore, we concentrate on computing- and communication-intensive systems with fluctuating demand for the infrastructure resources, where the use of cloud infrastructure is deemed highly suitable (Harms and Yamartino, 2010). Therefore, two cost components are considered:

- The costs of computing capacity, such as those incurred by hardware, software, and data storage; and
- Data communication costs.

These costs depend on whether the required capacity is acquired (private cloud) or utilized on a pay-per-use basis (public cloud); the cost of private cloud subsystems is constant whether the capacity is used or not, whereas the cost of public cloud depends on the volume of used capacity. If there are peaks in the demand for a resource and if this demand needs to be satisfied without a delay, then the use of the private cloud often leads to over-provisioning and to under-utilized resources.

Depending on the system's functionality and usage patterns, the adoption of a hybrid cloud may incur other costs, in addition to the costs of computing and data communication, such as, e.g., the cost of a load balancer responsible for intelligent division of load within the hybrid cloud, as well as the costs of persistent storage in the public cloud. However, the cost of the load balancer is assumed to be rather independent of the specific load division between the private and the public clouds, and hence this cost can be ignored when seeking cost-efficient division. On the other hand, some applications may require a significant volume of data to be persistently stored in a public cloud and thus may incur noticeable storage-related cost. The effect of such storage requirements on the cost depends on multiple factors and hence warrants a separate inquiry, which is left outside of the scope of this paper.

Similarly to (Weinman, 2011), the following assumptions are made:

1. Public cloud capacity is paid for only when used;
2. The other costs are either insignificant or do not depend on whether private or public cloud is used; and
3. The demand for the resources must be served without a delay.

In contrast to (Weinman, 2011), however, the data communication costs are not ignored in our model; it is assumed that the same pricing is applied for data uploading and downloading. As will be shown in the next section, the presence of

the data communication costs may have a significant effect on the overall costs and the optimal distribution between the private and the public cloud. Furthermore, whereas the unit cost of private and public cloud resources is assumed constant in (Weinman, 2011), in this work this assumption is relaxed – namely, it is assumed that the unit prices may change with the volume of the private or public capacity acquired.

3 Estimating the Costs of a Hybrid Cloud

In the previous section, the decomposition of the hybrid cloud solution into closed and open subsystems was introduced. In this section, based on the stated assumptions, the costs of open subsystems are derived.

3.1 Constituents of the costs of open subsystems

The costs of open subsystems are comprised of the costs of computing-related resources and the data communication costs, incurred both on the private and the public cloud sides:

$$C = C_{\mathrm{c}} + C_{\mathrm{b}}, \tag{1}$$

where

- C_{c} is the total cost of computing capacity (c) incurred;
- C_{b} is the cost of communication bandwidth (b) incurred.

These two costs can be decomposed into the costs incurred due to the private and public clouds:

$$C_{\mathrm{c}} = C_{\mathrm{co}} + C_{\mathrm{cc}}, \tag{2}$$

$$C_{\mathrm{b}} = C_{\mathrm{bo}} + C_{\mathrm{bc}}, \tag{3}$$

where

- C_{co} is the cost of computing capacity incurred with the private (o, own) cloud;
- C_{cc} is the cost of computing capacity incurred with the public cloud (c);
- C_{bo} is the cost of data communication incurred due to transferring the data to/from the private cloud;
- C_{bc} is the cost of data communication incurred due to transferring the data to/from the public cloud.

Thus,

$$C = C_{\mathrm{c}} + C_{\mathrm{b}} = C_{\mathrm{co}} + C_{\mathrm{cc}} + C_{\mathrm{bo}} + C_{\mathrm{bc}}. \tag{4}$$

Let p_{co}, p_{cc}, p_{bo}, and p_{bc} denote the price of a unit of the private cloud computing capacity, the public cloud computing capacity, the private cloud data communication capacity, and the public cloud data communication capacity, respectively.

Let us assume that, whenever a unit of computing capacity is demanded from the service, also k_{o} (k_{c}) units of data are transferred between the private (public) cloud and the customers of the service. Furthermore, let us assume that the volume of traffic transferred between the organization and the public cloud is proportional, with coefficient $0 < \rho < 1$, to the volume of the traffic between the public cloud and the customers (cf. Fig. 1). Having denoted the cumulative acquired private and public cloud computing capacity over time period T as L_{co} and L_{cc} respectively, it follows that:

- $L_{\mathrm{bo}} = k_{\mathrm{o}} L_{\mathrm{co}}$ of data is transferred between the private cloud and the customers;
- $L_{\mathrm{bc}} = k_{\mathrm{c}} L_{\mathrm{cc}}$ of data is transferred between the public cloud and the customers; and
- $L_{\mathrm{boc}} = \rho L_{\mathrm{bc}} = \rho k_{\mathrm{c}} L_{\mathrm{cc}}$ of data is transferred between the organization and the public cloud.

Individual costs can be evaluated as a product of the capacity volume and the unit price. The unit price is a function of volume $p(L)$, due to the quantity discounts, which will be discussed in the following subsection. Furthermore, it should be noted that:

- The unit price for the communication from/to the public cloud p_{bc} is determined by the volume of the data transferred both to/from the private cloud and to/from the customers;
- The unit price for the communication from/to the private cloud p_{bo} is determined by the volume of the data transferred both to/from the public cloud and to/from the customers.

Thus, the total costs can be rewritten in a form:

$$C = p_{\mathrm{co}} \times L_{\mathrm{co}} + p_{\mathrm{cc}} \times L_{\mathrm{cc}} + p_{\mathrm{bo}} \times (L_{\mathrm{bo}} + L_{\mathrm{boc}}) + p_{\mathrm{bc}} \times (L_{\mathrm{bc}} + L_{\mathrm{boc}}). \quad (5)$$

Since the total volume of data transferred to/from the private cloud is

$$L_{\mathrm{bo}} + L_{\mathrm{boc}} = k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}} \quad (6)$$

and since the total volume of data transferred to/from the public cloud is

$$L_{\mathrm{bc}} + L_{\mathrm{boc}} = k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}} \quad (7)$$

it follows that the costs in eq. (4) can be rewritten as:

$$C = p_{\mathrm{co}} L_{\mathrm{co}} + p_{\mathrm{cc}} L_{\mathrm{cc}} + p_{\mathrm{bo}}(k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}}) + p_{\mathrm{bc}}(k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}}). \quad (8)$$

In order to estimate the cost, both the prices and the volume of the acquired capacity need to be estimated. This is considered in the next two subsections.

3.2 Estimating the price of a unit of capacity

Often, pricing of a unit of capacity is not fixed but is instead a subject to market segmentation and price discrimination (Kotler and Keller, 2008). For instance, the first 10TB of Internet data traffic are charged by Amazon EC2, Ireland (Amazon Web Services, 2011) at the rate of 0.15 USD per GB, the next 40TB at the rate of 0.11 USD per GB, etc. This is a manifestation of the so called "second degree price discrimination" (Stole, 2003), where the unit price changes with the acquired quantity by means of all-units or incremental quantity discounts (Shah and Dixit, 2005; Schotanus et al, 2009).

Pricing in different segments can be assumed to follow a demand curve, whose shape is characterized by the price elasticity of demand. We will assume that the price elasticity of demand is constant, and hence the demand (manifested in the acquired capacity L) can be expressed as a function of the unit price (Perloff, 2008):

$$L = Ap^{\varepsilon}, \tag{9}$$

where A is a positive constant, and $\varepsilon \leq 0$ is the price elasticity of demand, assumed to be constant. Although the constant price elasticity of demand is unlikely to closely reflect the real pricing strategies, it is our belief that it allows the unit prices to be approximated more accurately than by assuming a fixed pricing.

From the equation above, the unit price can be expressed as a function of the acquired capacity:

$$p(L) = aL^{b}, \tag{10}$$

where $a = (1/A)^{-1/\varepsilon}$ is a positive constant and $b = 1/\varepsilon \leq 0$ represents an inverse of ε, i.e. the demand elasticity of price (Schotanus et al, 2009). The value of b determines how quickly the unit price declines with the acquired volume. Because of this and in order to avoid the confusion with the price elasticity of demand, we will refer to b, similarly to (Schotanus et al, 2009), as to the *steepness of the quantity discount*.

The values of a and b should be estimated over a period determined by the charging and billing rules of the service provider. For instance, Amazon price of 1GB of data transferred out of the EC2 depends on the monthly volume of the data transferred. Therefore, for Amazon EC2, the values of a and b should be estimated over a month.

Note that if $b < -1$ (i.e.$|\varepsilon| < 1$), then it would be economically more efficient for the customer to acquire (i.e. consume and be charged for) the maximum possible capacity, as the overall acquisition cost would be minimal:

$$C = Lp(L) = LaL^{b} = aL^{b+1}. \tag{11}$$

As could be seen, if $b < -1$, then the cost function above is a decreasing function of L; furthermore, for $L \to \infty$, it follows that $C \to 0$, which is unlikely to be realistic. Therefore, we will assume that the steepness of quantity demand is

less than 1 in absolute value ($|b| < 1$), corresponding to the so-called "relatively elastic" demand. Indeed, as will be considered later in the paper, for the data communication capacity in the public cloud, $b = -0.130$ (the estimate is based on (Amazon Web Services, 2011)); for the private computing capacity, $b = -0.478$ (the estimate is based on (Hamilton, 2010)). This is also in line with the real-world measurements (Bayoumi and Haacker, 2002) where the (absolute) price elasticity of demand for hardware was found to be in the range of $|\varepsilon| = 1.1 \ldots 1.8$, corresponding to $|b| = 0.56 \ldots 0.91$.

Given the constant price elasticity of demand – and hence the constant steepness of the quantity discount –, the unit prices of computing (p_{co} and p_{cc}) and data communication capacities (p_{bo} and p_{bc}) for the private and the public cloud respectively can be estimated as:

$$p_{\mathrm{co}} = a_{\mathrm{co}} L_{\mathrm{co}}^{b_{\mathrm{co}}}; \tag{12}$$

$$p_{\mathrm{cc}} = a_{\mathrm{cc}} L_{\mathrm{cc}}^{b_{\mathrm{cc}}}; \tag{13}$$

$$p_{\mathrm{bo}} = a_{\mathrm{bo}} (k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bo}}}; \tag{14}$$

$$p_{\mathrm{bc}} = a_{\mathrm{bc}} (k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bc}}}. \tag{15}$$

Then, the total costs of open subsystems can be rewritten as

$$\begin{aligned} C &= p_{\mathrm{co}} L_{\mathrm{co}} + p_{\mathrm{cc}} L_{\mathrm{cc}} + p_{\mathrm{bo}} (k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}}) + p_{\mathrm{bc}} (k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}}) \\ &= a_{\mathrm{co}} L_{\mathrm{co}}^{b_{\mathrm{co}}} L_{\mathrm{co}} + a_{\mathrm{cc}} L_{\mathrm{cc}}^{b_{\mathrm{cc}}} L_{\mathrm{cc}} + a_{\mathrm{bo}} (k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bo}}} (k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}}) \\ &+ a_{\mathrm{bc}} (k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bc}}} (k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}}), \end{aligned} \tag{16}$$

which can be simplified to:

$$C = a_{\mathrm{co}} L_{\mathrm{co}}^{b_{\mathrm{co}}+1} + a_{\mathrm{cc}} L_{\mathrm{cc}}^{b_{\mathrm{cc}}+1} + a_{\mathrm{bo}} (k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bo}}+1} + a_{\mathrm{bc}} (k_{\mathrm{c}} L_{\mathrm{cc}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bc}}+1} \tag{17}$$

or equally

$$C = a_{\mathrm{co}} L_{\mathrm{co}}^{b_{\mathrm{co}}+1} + a_{\mathrm{cc}} L_{\mathrm{cc}}^{b_{\mathrm{cc}}+1} + a_{\mathrm{bo}} (k_{\mathrm{o}} L_{\mathrm{co}} + \rho k_{\mathrm{c}} L_{\mathrm{cc}})^{b_{\mathrm{bo}}+1} + a_{\mathrm{bc}} [k_{\mathrm{c}} L_{\mathrm{cc}} (1+\rho)]^{b_{\mathrm{bc}}+1}. \tag{18}$$

Assuming for simplicity that the same software is used in both private and public open subsystems and that the demand is distributed between these subsystems independently of the expected data communication distribution, it follows that $k_{\mathrm{o}} = k_{\mathrm{c}} = k$, and hence the above can be rewritten as:

$$C = a_{\mathrm{co}} L_{\mathrm{co}}^{b_{\mathrm{co}}+1} + a_{\mathrm{cc}} L_{\mathrm{cc}}^{b_{\mathrm{cc}}+1} + a_{\mathrm{bo}} [k (L_{\mathrm{co}} + \rho L_{\mathrm{cc}})]^{b_{\mathrm{bo}}+1} + a_{\mathrm{bc}} [k L_{\mathrm{cc}} (1+\rho)]^{b_{\mathrm{bc}}+1}. \tag{19}$$

3.3 Estimating the acquired capacity

The estimation of acquired capacity differs for the public (L_{cc}) and the private (L_{co}) cloud.

For the *private* cloud, the acquired capacity can be treated as fixed whether or not it is used. Indeed, even if the private computing capacity is idle during a certain period of time, this capacity is still reserved for the purposes of service provisioning and hence incurs approximately the same costs as the actively used capacity would incur. This is due to the fact that the majority of costs factors, including the acquisition and integration costs, the costs of administration and maintenance, etc., are independent of the server load. This also applies to the data communication capacity, when the Internet Service Provider (ISP) charges for the bandwidth a fixed, bandwidth-dependent monthly fee – which is apparently the prevailing charging method used by ISPs (Stiller et al, 2001; Odlyzko, 2001). It should be noted that some of the costs, such as the costs of electricity, are affected by the server load, but the effect is not dramatic since the power consumption of an idle server still represents 65% of its peak consumption (Greenberg et al, 2008).

The capacity of the private cloud should be sufficient to serve the demand without a delay (assumption 3). Thus, in the private cloud, the acquired capacity can be estimated as the product of the maximum expected demand and the time. Let D denote the maximum demand for computing capacity observed over the estimation period T, and let q denote the threshold portion of that demand, up to which the demand is served with the private cloud. Then, the acquired private cloud computing capacity is:

$$L_{\mathrm{co}} = qDT. \tag{20}$$

For the *public* cloud, on the other hand, the acquired capacity represents the capacity used, and hence it depends on the characteristics of the demand curve. Therefore, in order to estimate L_{cc}, the demand curve needs to be analyzed.

Let us consider the demand curve $d(t)$, indicating how the demand for computing capacity changes with time. A realistic demand curve, which may have multiple peaks, *can be rearranged for the purpose of the analysis by sorting the data points in an ascending order, to make it a monotonically non-decreasing curve* (assumption 2 enables that), as shown in Fig. 2. Furthermore, for the sake of simplifying the analysis, let us assume that the rearranged demand curve is monotonically increasing.

Since the demand up to qD is served with the private cloud, the demand for the public cloud capacity is:

$$d_{\mathrm{c}}(t) = \begin{cases} 0, & \text{if } d(t) \leq qD; \\ d(t) - qD, & \text{otherwise.} \end{cases} \tag{21}$$

The acquired public cloud computing capacity can then be estimated as

$$L_{\mathrm{cc}} = \int_0^T d_{\mathrm{c}}(t)\ \mathrm{d}t = \int_{t_0}^T d_{\mathrm{c}}(t)\ \mathrm{d}t, \tag{22}$$

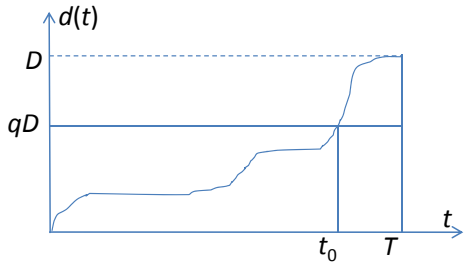

Fig. 2 Demand curve rearranged to be monotonically non-decreasing

and the equation (19) can be rewritten as

$$C = a_{\mathrm{co}}(qDT)^{b_{\mathrm{co}}+1} + a_{\mathrm{cc}}\left(\int_{t_0}^{T} d_{\mathrm{c}}(t)\ \mathrm{d}t\right)^{b_{\mathrm{cc}}+1} + a_{\mathrm{bo}}\left[k\left(qDT + \rho\int_{t_0}^{T} d_{\mathrm{c}}(t)\ \mathrm{d}t\right)\right]^{b_{\mathrm{bo}}+1} + a_{\mathrm{bc}}\left[k(1+\rho)\int_{t_0}^{T} d_{\mathrm{c}}(t)\ \mathrm{d}t\right]^{b_{\mathrm{bc}}+1}. \quad (23)$$

The cost can be seen as a function of the private cloud threshold q. In a cost-efficient hybrid cloud the threshold portion of the private cloud equals $q_{\min} = \min_q C$. In the next section, we will consider how the value of $q_{\min}$ depends on other variables.

4 Analyzing the Hybrid Cloud Costs

In this section, we will analyze how the value of $q_{\min}$ minimizing the cost of the open subsystems depends on other variables. First, the case of fixed unit prices for computing and data communication capacity is considered. After that, the effect of quantity discounting is analyzed.

4.1 Fixed unit prices

According to eq. (8) and (19) above, the cost of open subsystems is a function of the acquired computing capacity, both in the private and in the public cloud, which in turn depends on the distribution of the capacity between the private and the public cloud, as regulated by the value of q. Furthermore, the open subsystem cost depends on i) how intensive the communication that occurs between the system and its customers is, as reflected in the value of k; and on ii) how intensive the interaction that is needed between the private and public subsystems is, as reflected in the value of ρ.

Here, we consider the case when the unit price of capacity is fixed, i.e. the effect of quantity discounting can be ignored: $b_{\mathrm{bo}} = b_{\mathrm{bc}} = b_{\mathrm{co}} = b_{\mathrm{cc}} = 0$. Thus, a_{bo}, a_{bc}, a_{co}, and a_{cc} represent the fixed unit prices of acquired capacities.

We will also assume that the unit price of capacity in the private cloud is less or equally expensive compared to the unit price in the public clouds, in line with the findings of Khajeh-Hosseini et al. (Khajeh-Hosseini et al, 2011). The higher unit price of a public cloud can be partly attributed to the margins added by the cloud provider on top of its costs. Therefore,

$$a_{\mathrm{cc}} = u_{\mathrm{c}} a_{\mathrm{co}}, \tag{24}$$

$$a_{\mathrm{bc}} = u_{\mathrm{b}} a_{\mathrm{bo}}, \tag{25}$$

where $u_{\mathrm{c}} \geq 1$ and $u_{\mathrm{b}} \geq 1$. Now the expression for the cost of open subsystems can be rewritten as:

$$C = a_{\mathrm{co}} qDT + u_{\mathrm{c}} a_{\mathrm{co}} L_{\mathrm{cc}} + a_{\mathrm{bo}} k(qDT + \rho L_{\mathrm{cc}}) + u_{\mathrm{b}} a_{\mathrm{bo}} k(1+\rho) L_{\mathrm{cc}}. \tag{26}$$

The unit prices a_{co} and a_{bo}, as well as u_{c} and u_{b}, can be seen as constants whose values are estimated by consulting public cloud providers' price lists (for the public cloud) or by estimating the acquisition and operations costs over the depreciation period (for the private cloud).

Proposition 1 *The cost of open subsystems in the hybrid cloud increases, as the data communication intensity grows.*

Proof The correctness of this proposition can be easily shown by taking partial derivatives of C with respect to k and ρ, which reflect the data communication intensity of the service. Based on eq. (19) for the open subsystem costs, it can be shown that (note that the assumptions on fixed prices and on public cloud capacity being more expensive are not needed for the proof):

$$\frac{\partial C}{\partial k} = a_{\mathrm{bo}}(L_{\mathrm{co}} + \rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)k^{b_{\mathrm{bo}}} + a_{\mathrm{bc}}[L_{\mathrm{cc}}(1+\rho)]^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)k^{b_{\mathrm{bc}}} > 0; \tag{27}$$

$$\begin{aligned}\frac{\partial C}{\partial \rho} = {} & a_{\mathrm{bo}} k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}} + \rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}} L_{\mathrm{cc}} \\ & + a_{\mathrm{bc}}(kL_{\mathrm{cc}})^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)(1+\rho)^{b_{\mathrm{bc}}+1} > 0.\end{aligned} \tag{28}$$

The positivity of $\frac{\partial C}{\partial k}$ and $\frac{\partial C}{\partial \rho}$ follows from the positivity of all the variables except b_{bo} and b_{bc}. Furthermore, since $|b| < 1$, it follows that $b_{\mathrm{bo}} + 1$ and $b_{\mathrm{bc}} + 1$ are positive values, and hence their respective terms are positive, too.

Thus, provided the price of data communication capacity is non-zero ($p_{\mathrm{bo}} > 0$), and provided that at least some of the capacity is acquired from the public cloud ($t_0 < T$ and hence $L_{\mathrm{cc}} = \int_{t_0}^{T} d_{\mathrm{c}}(t)\ \mathrm{d}t > 0$), the values of the partial derivatives in (27) and (28) are positive. Therefore, the costs increase as k and ρ values grow, q.e.d. □

If only the private cloud capacity is used, then $t_0 = T$ and hence $\frac{\partial C}{\partial \rho} = 0$. This reflects the fact that no data communication between the organization and the public cloud takes place, and hence such communication has no effect on the open subsystem costs.

Proposition 2 *If $u_c > 1$ and $u_b > 1$, then a hybrid cloud has lower costs than a fully private cloud or fully public cloud solution.*

Proof Let us find the value of q that minimizes the costs of the open subsystems.

Consider the open subsystem costs. Eq. (26) can be rewritten as:

$$\begin{aligned} C &= a_{co} q DT + u_c a_{co} L_{cc} + a_{bo} k (qDT + \rho L_{cc}) + u_b a_{bo} k (1+\rho) L_{cc} \\ &= (a_{co} + a_{bo} k) qDT + [u_c a_{co} + a_{bo} k \rho + u_b a_{bo} k (1+\rho)] L_{cc} \\ &= (a_{co} + a_{bo} k) qDT + [u_c a_{co} + a_{bo} k (\rho + u_b + \rho u_b)] L_{cc} \\ &= (a_{co} + a_{bo} k) qDT + [u_c a_{co} + a_{bo} k (u_b + \rho(1+u_b))] L_{cc}. \end{aligned} \tag{29}$$

The partial derivative of C with respect to q is:

$$\frac{\partial C}{\partial q} = (a_{co} + a_{bo} k) DT + [u_c a_{co} + a_{bo} k (u_b + \rho(1+u_b))] \frac{\partial}{\partial q} L_{cc}. \tag{30}$$

Let $t_c(d)$ denote the inverse function of $d(t)$. Let us also define function $\tau_c(d)$:

$$\tau_c(d) = T - t_c(d). \tag{31}$$

The value of $\tau_c(d_0)$, where $d_0 = qD$, indicates the amount of time during which the public cloud capacity is used given the value of q. Then, the acquired public cloud computing capacity L_{cc} can be evaluated by integrating over d:

$$L_{cc} = \int_{d_0}^{D} \tau_c(d) \ \mathrm{d}d. \tag{32}$$

Let $F(d)$ be an anti-derivative of $\tau_c(d)$. Then,

$$L_{cc} = \int_{d_0}^{D} \tau_c(d) \ \mathrm{d}d = F(D) - F(d_0). \tag{33}$$

Note that $F(D)$ is independent of q, whereas $F(d_0)$ depends on q, since d_0 is a function on q. Therefore,

$$\begin{aligned} \frac{\partial}{\partial q} L_{cc} &= \frac{\partial}{\partial q} \left(\int_{d_0}^{D} \tau_c(d) \ \mathrm{d}d \right) = \frac{\partial}{\partial q} F(D) - \frac{\partial}{\partial q} F(d_0) \\ &= -\frac{\partial}{\partial q} F(d_0) = -\frac{\partial F(d_0)}{\partial d} \frac{\partial d}{\partial q} = -\tau_c(d_0) D. \end{aligned} \tag{34}$$

Then,

$$\frac{\partial C}{\partial q} = (a_{co} + a_{bo} k) DT - [u_c a_{co} + a_{bo} k (u_b + \rho(1+u_b))] \tau_c(d_0) D. \tag{35}$$

The second derivative is:

$$\begin{aligned}\frac{\partial^2 C}{\partial q^2} &= -[u_c a_{co} + a_{bo} k(u_b + \rho(1+u_b))] D \frac{\partial}{\partial q}\tau_c(d_0) \\ &= -[u_c a_{co} + a_{bo} k(u_b + \rho(1+u_b))] D \frac{\partial \tau_c(d_0)}{\partial d}\frac{\partial d}{\partial q} \\ &= -[u_c a_{co} + a_{bo} k(u_b + \rho(1+u_b))] D^2 \frac{\partial \tau_c(d_0)}{\partial d}. \end{aligned} \quad (36)$$

Recall that $t_c(d)$ is inverse function of $d(t)$; furthermore, $d(t)$ is monotonically increasing. According to the inverse function theorem, for a domain where $d(t)$ is increasing, it holds that

$$\frac{\partial}{\partial d} t_c(d) = \frac{1}{\frac{\partial}{\partial t} d(t)}. \quad (37)$$

Since $d(t)$ is increasing in this domain, it follows that $\frac{\partial}{\partial t} d(t) > 0$, and hence $\frac{\partial}{\partial d} t_c(d) > 0$. From here, we get:

$$\frac{\partial}{\partial d}\tau_c(d_0) = \frac{\partial}{\partial d}(T - t_c(d)) = -\frac{\partial}{\partial d} t_c(d) < 0. \quad (38)$$

Thus, it follows that the second derivative is positive:

$$\frac{\partial^2 C}{\partial q^2} > 0. \quad (39)$$

Since $\frac{\partial^2 C}{\partial q^2}$ is positive, it follows that, if there is a value of $q_{\min} \in [0,1]$ such that the first derivative $\frac{\partial C(q_{\min})}{\partial q} = 0$, then $q_{\min}$ minimizes C, i.e.

$$\frac{\partial C(q_{\min})}{\partial q} = (a_{co} + a_{bo} k) DT - [u_c a_{co} + a_{bo} k(u_b + \rho(1+u_b))]\tau_c(d_0) D = 0. \quad (40)$$

Observing that τ_c is also a function of q we obtain:

$$\tau_c(d_0, q_{\min}) = \frac{(a_{co} + a_{bo} k)T}{u_c a_{co} + a_{bo} k(u_b + \rho(1+u_b))}. \quad (41)$$

By solving eq. (41), the value of $q_{\min}$ can be found. Since $u_c > 1$, $u_b > 1$, and since the unit prices are positive, it follows that

$$0 < \frac{a_{co} + a_{bo} k}{u_c a_{co} + a_{bo} k(u_b + \rho(1+u_b))} < 1,$$

and hence $0 < \tau_c(d_0, q_{\min}) < T$.

Recall that the values of $\tau_c(q)$ are within the region $[0, T]$. Furthermore, from eq. (38) it follows that $\tau_c(q)$ is monotonically decreasing function in the domain $(0, 1)$. Therefore, there exists a value $q_{\min} \in (0, 1)$ satisfying eq. (41), i.e. a hybrid solution has lower costs than a purely private cloud ($q = 1$) or purely public cloud ($q = 0$) solution, q.e.d. □

Corollary 1 *In the absence of data communication costs ($k = 0$), the portion of the time when public cloud is used should be the inverse of the premium charged by the cloud software vendor*

Given $k = 0$, eq. (41) can be rewritten as:

$$\frac{\tau_c(d_0, q_{\min})}{T} = \frac{a_{co}}{u_c a_{co}} = \frac{1}{u_c}. \tag{42}$$

This is in line with (Weinman, 2011) where it was shown that in the absence of data communication costs, and for the uniformly distributed demand, the cost-optimal portion of public cloud capacity (i.e. $1 - q_{\min}$) is the inverse of u_c. Indeed, for the uniformly distributed demand,

$$\tau_c(d_0, q_{\min}) = T(1 - q_{\min}). \tag{43}$$

If $k = 0$, then eq. (41) simplifies to

$$T(1 - q_{\min}) = \frac{a_{co} T}{u_c a_{co}} \tag{44}$$

It follows that $1 - q_{\min} = 1/u_c$, as in (Weinman, 2011). Note that, according to this corollary, the regularity represented by eq. (42) holds for the generic case of arbitrary monotonically increasing demand function, whereas only a special case of uniformly distributed demand was considered in (Weinman, 2011).

Proposition 3 *If $u_b \approx u_c$, then the greater the data communication intensity of the service, as indicated by k and ρ, the more private cloud capacity is needed to minimize the costs.*

Proof Let $Q(\tau_c)$ be the inverse function of $\tau_c(q)$, i.e.

$$q = Q(\tau_c). \tag{45}$$

Recall that from eq. (41) the value of q minimizing C can be found. By substituting (45) into eq. (41) we can express the value of $q_{\min}$ as

$$q_{\min} = Q(\tau_c) = Q\left(\frac{(a_{co} + a_{bo}k)T}{u_c a_{co} + a_{bo}k(u_b + \rho(1 + u_b))}\right). \tag{46}$$

Let us consider how $q_{\min}$ (and hence Q) depends on k. Using the chain rule:

$$\frac{\partial Q}{\partial k} = \frac{\partial Q}{\partial \tau_c}\frac{\partial \tau_c}{\partial k}. \tag{47}$$

By using the inverse function theorem, and applying the chain rule, we obtain

$$\frac{\partial Q}{\partial \tau_c} = \frac{1}{\frac{\partial \tau_c}{\partial q}} = \frac{1}{\frac{\partial \tau_c}{\partial d}\frac{\partial d}{\partial q}}. \tag{48}$$

Since $\frac{\partial \tau_c}{\partial d} < 0$ (according to 38) and since $\frac{\partial d}{\partial q} = D$, it follows that $\frac{\partial Q}{\partial \tau_c} < 0$.

By taking partial derivatives from both sides of eq. (41), we obtain:

$$\begin{aligned}\frac{\partial \tau_c}{\partial k} &= \frac{\partial}{\partial k}\left(\frac{(a_{co} + a_{bo}k)T}{u_c a_{co} + a_{bo}k(u_b + \rho(1 + u_b))}\right) \\ &= -a_{bo}a_{co}T\frac{\rho(1 + u_b) + u_b - u_c}{[u_c a_{co} + a_{bo}k(u_b + \rho(1 + u_b))]^2} < 0.\end{aligned} \tag{49}$$

Given $u_{\mathrm{b}} \approx u_{\mathrm{c}}$, the term $\rho(1+u_{\mathrm{b}})+u_{\mathrm{b}}-u_{\mathrm{c}}$ is positive and hence $\frac{\partial \tau_{\mathrm{c}}}{\partial k}$ is negative.

Thus, $\frac{\partial Q}{\partial \tau_{\mathrm{c}}} < 0$ and $\frac{\partial \tau_{\mathrm{c}}}{\partial k} < 0$. Since both terms in the RHS of (47) are negative, their product is positive, i.e. $\frac{\partial Q}{\partial k} > 0$, implying that $q_{\min}$ increases as k grows.

Similarly, the dependence of $q_{\min}$ (and hence Q) on ρ can be investigated. Using the chain rule:

$$\frac{\partial Q}{\partial \rho} = \frac{\partial Q}{\partial \tau_{\mathrm{c}}} \frac{\partial \tau_{\mathrm{c}}}{\partial \rho}. \tag{50}$$

By taking partial derivatives from both sides of eq. (41), we obtain:

$$\begin{aligned} \frac{\partial \tau_{\mathrm{c}}}{\partial \rho} &= \frac{\partial}{\partial \rho} \left(\frac{(a_{\mathrm{co}} + a_{\mathrm{bo}} k)T}{u_{\mathrm{c}} a_{\mathrm{co}} + a_{\mathrm{bo}} k(u_{\mathrm{b}} + \rho(1+u_{\mathrm{b}}))} \right) \\ &= -\frac{(a_{\mathrm{co}} + a_{\mathrm{bo}} k)T a_{\mathrm{bo}} k(1+u_{\mathrm{b}})}{[u_{\mathrm{c}} a_{\mathrm{co}} + a_{\mathrm{bo}} k(u_{\mathrm{b}} + \rho(1+u_{\mathrm{b}}))]^2} < 0. \end{aligned} \tag{51}$$

Thus, $\frac{\partial Q}{\partial \tau_{\mathrm{c}}} < 0$ and $\frac{\partial \tau_{\mathrm{c}}}{\partial \rho} < 0$. Since both terms in the RHS of (50) are negative, their product is positive, i.e. $\frac{\partial Q}{\partial \rho} > 0$. Hence, $q_{\min}$ increases as ρ grows.

Above, it has been shown that $q_{\min}$ increases with either k or ρ. This suggests that the greater the values of k or ρ the greater the portion of the capacity that should be allocated to the private cloud, q.e.d. □

Corollary 2 *In the special case of $u_{\mathrm{c}} \gg u_{\mathrm{b}}$, the greater the data communication intensity, as indicated by k, the less the amount of private computing capacity that should be acquired.*

It can be seen that when $u_{\mathrm{c}} \gg u_{\mathrm{b}}$ the partial derivative $\frac{\partial \tau_{\mathrm{c}}}{\partial k}$ becomes positive. As a result, according to (47), $\frac{\partial Q}{\partial k} < 0$ and hence $q_{\min}$ decreases as k grows.

It should be noted that the special case of $u_{\mathrm{c}} \gg u_{\mathrm{b}}$ does not change the effect of ρ, i.e. greater values of ρ lead to an increase in the value of $q_{\min}$, even if $u_{\mathrm{c}} \gg u_{\mathrm{b}}$.

4.2 Effect of quantity discounting

Here, the combined effect of i) the form of the demand function, and ii) quantity discounting on the value of $q_{\min}$ is analyzed. First, however, for the sake of illustrating the effect of quantity discounting on optimal $q_{\min}$ in a hybrid cloud, let us consider the case of constant demand function, i.e.:

$$d(t) = D. \tag{52}$$

Given the constant demand function above, it follows that $L_{\mathrm{cc}} = (1-q)DT$.

Proposition 4 *If the demand is constant, the minimum of costs occurs for $q = 0$ or $q = 1$, i.e. a hybrid cloud is more expensive than a private or a public cloud.*

Proof Assuming for simplicity that $b_{\mathrm{co}} = b_{\mathrm{cc}} = b_{\mathrm{bo}} = b_{\mathrm{bc}} = b$, and given constant demand, the cost function in eq. (19) can be rewritten as:

$$C = a_{\mathrm{co}}(qDT)^{b+1} + a_{\mathrm{cc}}((1-q)DT)^{b+1} + a_{\mathrm{bo}}[k(qDT + \rho(1-q)DT)]^{b+1} + a_{\mathrm{bc}}[k(1-q)DT(1+\rho)]^{b+1}, \quad (53)$$

which can be further re-grouped as

$$C = (DT)^{b+1}\big[a_{\mathrm{co}}q^{b+1} + a_{\mathrm{cc}}(1-q)^{b+1} + a_{\mathrm{bo}}[k(q+\rho(1-q))]^{b+1} + a_{\mathrm{bc}}[k(1-q)(1+\rho)]^{b+1}\big]. \quad (54)$$

The partial derivative of C with respect to q is:

$$\frac{\partial C}{\partial q} = (DT)^{b+1}\big[a_{\mathrm{co}}(b+1)q^{b} - a_{\mathrm{cc}}(b+1)(1-q)^{b} + a_{\mathrm{bo}}k^{b+1}(b+1)(q+\rho(1-q))^{b}(1-\rho) - a_{\mathrm{bc}}[k(1+\rho)]^{b+1}(b+1)(1-q)^{b}\big], \quad (55)$$

which can be further re-grouped as:

$$\frac{\partial C}{\partial q} = (DT)^{b+1}(b+1)\Big[a_{\mathrm{co}}q^{b} - a_{\mathrm{cc}}(1-q)^{b} + a_{\mathrm{bo}}k^{b+1}(q+\rho(1-q))^{b}(1-\rho) - a_{\mathrm{bc}}[k(1+\rho)]^{b+1}(1-q)^{b}\Big]. \quad (56)$$

The second derivative with respect to q takes the form:

$$\frac{\partial^2 C}{\partial q^2} = (DT)^{b+1}(b+1)\big[a_{\mathrm{co}}bq^{b-1} + a_{\mathrm{cc}}b(1-q)^{b-1} + a_{\mathrm{bo}}k^{b+1}b(q+\rho(1-q))^{b-1}(1-\rho)^2 + a_{\mathrm{bc}}[k(1+\rho)]^{b+1}b(1-q)^{b-1}\big] \quad (57)$$

or equally

$$\frac{\partial^2 C}{\partial q^2} = (DT)^{b+1}(b+1)b\big[a_{\mathrm{co}}q^{b-1} + a_{\mathrm{cc}}(1-q)^{b-1} + a_{\mathrm{bo}}k^{b+1}(q+\rho(1-q))^{b-1}(1-\rho)^2 + a_{\mathrm{bc}}[k(1+\rho)]^{b+1}(1-q)^{b-1}\big]. \quad (58)$$

Since $b < 0$ and $|b| < 1$, it follows that $\frac{\partial^2 C}{\partial q^2} < 0$ and hence the cost function is concave. Hence, the minimum occurs at an edge ($q = 0$ or $q = 1$), q.e.d. □

Corollary 3 *If the demand is flat, and it holds that $a_{\mathrm{cc}} \geq a_{\mathrm{co}}$ and $a_{\mathrm{bc}} \approx a_{\mathrm{bo}}$, then the costs are at minimum when $q = 1$, i.e. the private cloud deployment provides the minimum costs.*

Proof Consider the edge values of q:

$$q = 0: \quad C = (DT)^{b+1}\left[a_{\mathrm{cc}} + a_{\mathrm{bo}}k^{b+1}\rho^{b+1} + a_{\mathrm{bc}}[k(1+\rho)]^{b+1}\right], \tag{59}$$

$$q = 1: \quad C = (DT)^{b+1}\left[a_{\mathrm{co}} + a_{\mathrm{bo}}k^{b+1}\right]. \tag{60}$$

Let us compare the costs for $q = 0$ and $q = 1$. If the costs are greater for $q = 0$, then:

$$a_{\mathrm{cc}} + a_{\mathrm{bo}}k^{b+1}\rho^{b+1} + a_{\mathrm{bc}}[k(1+\rho)]^{b+1} > a_{\mathrm{co}} + a_{\mathrm{bo}}k^{b+1}. \tag{61}$$

Since $a_{\mathrm{cc}} \geq a_{\mathrm{co}}$, it follows that $a_{\mathrm{cc}} = ua_{\mathrm{co}}$, where $u \geq 1$. Recalling that $a_{\mathrm{bc}} \approx a_{\mathrm{bo}}$, it follows that:

$$ua_{\mathrm{co}} + a_{\mathrm{bo}}k^{b+1}\rho^{b+1} + a_{\mathrm{bo}}[k(1+\rho)]^{b+1} > a_{\mathrm{co}} + a_{\mathrm{bo}}k^{b+1}; \tag{62}$$

$$a_{\mathrm{co}}(u-1) + a_{\mathrm{bo}}k^{b+1}\rho^{b+1} + a_{\mathrm{bo}}[k(1+\rho)]^{b+1} > a_{\mathrm{bo}}k^{b+1}; \tag{63}$$

$$a_{\mathrm{co}}(u-1) + a_{\mathrm{bo}}k^{b+1}(\rho^{b+1} + (1+\rho)^{b+1}) > a_{\mathrm{bo}}k^{b+1}; \tag{64}$$

$$a_{\mathrm{co}}(u-1) + a_{\mathrm{bo}}k^{b+1}(\rho^{b+1} + (1+\rho)^{b+1} - 1) > 0. \tag{65}$$

Since $\rho \geq 0$ and $u \geq 1$, the inequality above always holds, q.e.d. □

Thus, in case the demand is constant and the unit price of computing capacity is greater in the public cloud, while the unit prices of data-communication capacity are approximately equal, the use of private cloud is cost-efficient.

Corollary 4 *In communication-intensive services with a flat demand, the costs are at minimum when $q = 1$, i.e. private cloud deployment provides the minimum costs.*

Proof Indeed, if k and/or ρ is large, then $[k(1+\rho)]^{b+1}$ is large and hence the condition in eq. (61) holds. Thus, $\frac{\partial C}{\partial q} < 0$, suggesting that the cost decreases as q increases and hence the minimum occurs when $q = 1$, q.e.d. □

The proposition above illustrated the effect of quantity discounting in the case of a constant demand function. Now, let us return to the case of an arbitrary demand distribution (see eq. (19)):

$$C = a_{\mathrm{co}}L_{\mathrm{co}}^{b_{\mathrm{co}}+1} + a_{\mathrm{cc}}L_{\mathrm{cc}}^{b_{\mathrm{cc}}+1} + a_{\mathrm{bo}}[k(L_{\mathrm{co}} + \rho L_{\mathrm{cc}})]^{b_{\mathrm{bo}}+1} + a_{\mathrm{bc}}[kL_{\mathrm{cc}}(1+\rho)]^{b_{\mathrm{bc}}+1}.$$

Proposition 5 *In case the unit price of capacity is subject to a quantity discount, either private or public cloud deployment, but not a hybrid cloud, may provide minimal costs.*

Proof (for a special case only) Consider the partial derivative of C with respect to q:

$$\begin{aligned}\frac{\partial C}{\partial q} &= a_{\mathrm{co}}(b_{\mathrm{co}}+1)L_{\mathrm{co}}^{b_{\mathrm{co}}}\frac{\partial}{\partial q}L_{\mathrm{co}} + a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}}\frac{\partial}{\partial q}L_{\mathrm{cc}} \\ &\quad + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}} + \rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\left(\frac{\partial}{\partial q}L_{\mathrm{co}} + \rho\frac{\partial}{\partial q}L_{\mathrm{co}}\right) \\ &\quad + a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\frac{\partial}{\partial q}L_{\mathrm{cc}},\end{aligned} \tag{66}$$

which can be rewritten in the form

$$\begin{aligned}\frac{\partial C}{\partial q} &= \left[a_{\mathrm{co}}(b_{\mathrm{co}}+1)L_{\mathrm{co}}^{b_{\mathrm{co}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\right]\frac{\partial}{\partial q}L_{\mathrm{co}} \\ &+ \left[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\rho\right. \\ &+ \left.a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\right]\frac{\partial}{\partial q}L_{\mathrm{cc}}. \end{aligned} \tag{67}$$

The second derivative is:

$$\begin{aligned}\frac{\partial^2 C}{\partial q^2} &= \frac{\partial}{\partial q}\left(a_{\mathrm{co}}(b_{\mathrm{co}}+1)L_{\mathrm{co}}^{b_{\mathrm{co}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\right)\frac{\partial}{\partial q}L_{\mathrm{co}} \\ &+ \left[a_{\mathrm{co}}(b_{\mathrm{co}}+1)L_{\mathrm{co}}^{b_{\mathrm{co}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\right]\frac{\partial^2}{\partial q^2}L_{\mathrm{co}} \\ &+ \frac{\partial}{\partial q}\left[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\rho\right. \\ &+ \left.a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\right]\frac{\partial}{\partial q}L_{\mathrm{cc}} \\ &+ \left[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\rho\right. \\ &+ \left.a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\right]\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}. \end{aligned} \tag{68}$$

Note that $\frac{\partial}{\partial q}L_{\mathrm{co}} = DT$ and hence $\frac{\partial^2}{\partial q^2}L_{\mathrm{co}} = 0$, i.e. the above expression is simplified to:

$$\begin{aligned}\frac{\partial^2 C}{\partial q^2} &= \frac{\partial}{\partial q}\left(a_{\mathrm{co}}(b_{\mathrm{co}}+1)L_{\mathrm{co}}^{b_{\mathrm{co}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\right)\frac{\partial}{\partial q}L_{\mathrm{co}} \\ &+ \frac{\partial}{\partial q}\left[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\rho\right. \\ &+ \left.a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\right]\frac{\partial}{\partial q}L_{\mathrm{cc}} \\ &+ \left[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}} + a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\rho\right. \\ &+ \left.a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\right]\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}. \end{aligned} \tag{69}$$

which, by opening the partial derivatives, can be rewritten in a form:

$$\begin{aligned}\frac{\partial^2 C}{\partial q^2} = &\Bigg(a_{\mathrm{co}}(b_{\mathrm{co}}+1)b_{\mathrm{co}}L_{\mathrm{co}}^{b_{\mathrm{co}}-1}\frac{\partial}{\partial q}L_{\mathrm{co}}\\ &+ a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)b_{\mathrm{bo}}(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}-1}\left(\frac{\partial}{\partial q}L_{\mathrm{co}}+\rho\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)\Bigg)\frac{\partial}{\partial q}L_{\mathrm{co}}\\ &+\Bigg[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)b_{\mathrm{cc}}L_{\mathrm{cc}}^{b_{\mathrm{cc}}-1}\frac{\partial}{\partial q}L_{\mathrm{cc}}\\ &+ a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}\rho(b_{\mathrm{bo}}+1)b_{\mathrm{bo}}(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}-1}\left(\frac{\partial}{\partial q}L_{\mathrm{co}}+\rho\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)\\ &+ a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)b_{\mathrm{bc}}L_{\mathrm{cc}}^{b_{\mathrm{bc}}-1}\frac{\partial}{\partial q}L_{\mathrm{cc}}\Bigg]\frac{\partial}{\partial q}L_{\mathrm{cc}}\\ &+\Bigg[a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}}+a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}}\rho\\ &+ a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}}\Bigg]\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}. \end{aligned} \tag{70}$$

Finally, the expression above can be regrouped in a form:

$$\begin{aligned}\frac{\partial^2 C}{\partial q^2} = &\, a_{\mathrm{co}}(b_{\mathrm{co}}+1)b_{\mathrm{co}}L_{\mathrm{co}}^{b_{\mathrm{co}}-1}\left(\frac{\partial}{\partial q}L_{\mathrm{co}}\right)^2\\ &+ a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}-1}\times\left[b_{\mathrm{cc}}\left(\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2+L_{\mathrm{cc}}\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}\right]\\ &+ a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}-1}\times\left[b_{\mathrm{bc}}\left(\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2+L_{\mathrm{cc}}\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}\right]\\ &+ a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}-1}\\ &\times\left[b_{\mathrm{bo}}\left(\frac{\partial}{\partial q}L_{\mathrm{co}}+\rho\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2+\rho(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}\right]\end{aligned}$$

and further rewritten as

$$\begin{aligned}\frac{\partial^2 C}{\partial q^2} = &\, B1 + a_{\mathrm{cc}}(b_{\mathrm{cc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{cc}}-1}\times B2\\ &+ a_{\mathrm{bc}}k^{b_{\mathrm{bc}}+1}(1+\rho)^{b_{\mathrm{bc}}+1}(b_{\mathrm{bc}}+1)L_{\mathrm{cc}}^{b_{\mathrm{bc}}-1}\times B3\\ &+ a_{\mathrm{bo}}k^{b_{\mathrm{bo}}+1}(b_{\mathrm{bo}}+1)(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})^{b_{\mathrm{bo}}-1}\times B4,\end{aligned} \tag{71}$$

where the terms $B1$, $B2$, $B3$ and $B4$ correspond, respectively, to:

$$B1 = a_{\mathrm{co}}(b_{\mathrm{co}}+1)b_{\mathrm{co}}L_{\mathrm{co}}^{b_{\mathrm{co}}-1}\left(\frac{\partial}{\partial q}L_{\mathrm{co}}\right)^2; \tag{72}$$

$$B2 = b_{\mathrm{cc}}\left(\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2+L_{\mathrm{cc}}\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}; \tag{73}$$

$$B3 = b_{\mathrm{bc}}\left(\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2+L_{\mathrm{cc}}\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}; \tag{74}$$

$$B4 = b_{\mathrm{bo}}\left(\frac{\partial}{\partial q}L_{\mathrm{co}}+\rho\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2+\rho(L_{\mathrm{co}}+\rho L_{\mathrm{cc}})\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}}. \tag{75}$$

The sign of $\frac{\partial^2 C}{\partial q^2}$ depends on the signs of the four constituents in eq. (71). As could be seen, the first constituent is always negative ($B1 < 0$), whereas the signs of the other three constituents are determined by the signs of the terms $B2$, $B3$ and $B4$.

Having observed that:

$$\frac{\partial}{\partial q} L_{\mathrm{cc}} = -\tau_{\mathrm{c}}(d_0) D < 0;$$
$$\frac{\partial}{\partial d} \tau_{\mathrm{c}}(d_0) = \frac{\partial}{\partial d}(T - t_{\mathrm{c}}(d_0)) = -\frac{\partial}{\partial d} t_{\mathrm{c}}(d_0) < 0,$$

it follows that

$$\frac{\partial^2}{\partial q^2} L_{\mathrm{cc}} = -D \frac{\partial}{\partial q} \tau_{\mathrm{c}}(d_0) = -D \frac{\partial \tau_{\mathrm{c}}(d_0)}{\partial d} \frac{\partial d}{\partial q} = -D^2 \frac{\partial \tau_{\mathrm{c}}(d_0)}{\partial d} > 0. \tag{76}$$

Based on the above observations, it can be seen that the signs of $B2$, $B3$ and $B4$ depend

- on the absolute values of the steepness of the quantity discount (b_{cc}, b_{bc}, b_{bo}), on one hand, and
- on the particular form of the acquired capacity function L_{cc}, on the other hand (the sign of $B4$ is also affected by ρ).

Let us consider separately the cases of small and large absolute values of the steepness:

1. For small absolute values of b_{co}, b_{cc}, b_{bc}, and b_{bo}, the term $B1 \to 0$ while the terms $B2$, $B3$ and $B4$ are positive, and hence $\frac{\partial^2 C}{\partial q^2}$ is positive:

$$b_{\mathrm{cc}} \to 0; \text{ therefore } B2 \to L_{\mathrm{cc}} \frac{\partial^2}{\partial q^2} L_{\mathrm{cc}} > 0;$$
$$b_{\mathrm{bc}} \to 0; \text{ therefore } B3 \to L_{\mathrm{cc}} \frac{\partial^2}{\partial q^2} L_{\mathrm{cc}} > 0;$$
$$b_{\mathrm{bo}} \to 0; \text{ therefore } B4 \to (L_{\mathrm{co}} + \rho L_{\mathrm{cc}}) \frac{\partial^2}{\partial q^2} L_{\mathrm{cc}} > 0.$$

   In fact, when the steepness of the quantity discount is small ($b \to 0$), expression (71) simplifies to the case considered in the preceding section, namely:

$$\frac{\partial^2 C}{\partial q^2} = [a_{\mathrm{cc}} + a_{\mathrm{bo}} k \rho + a_{\mathrm{bc}} k (1 + \rho)] \frac{\partial^2}{\partial q^2} L_{\mathrm{cc}} > 0. \tag{77}$$

   thus implying that the cost is a convex function of q.
2. As the absolute values of the coefficients b_{cc}, b_{bc}, and b_{bo} increase, the signs of the terms $B2$, $B3$ and $B4$ change from positive to negative, and hence $\frac{\partial^2 C}{\partial q^2}$ is

becoming negative. Let us demonstrate it for the linear demand distribution curve. In this case, $L_{\mathrm{cc}} = \frac{1}{2}DT(1-q)^2$, and hence:

$$\frac{\partial}{\partial q}L_{\mathrm{co}} = DT; \tag{78}$$

$$\frac{\partial}{\partial q}L_{\mathrm{cc}} = -DT(1-q); \tag{79}$$

$$\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}} = DT. \tag{80}$$

Thus, for $B2$, we obtain:

$$\begin{aligned} B2 &= b_{\mathrm{cc}}\left(\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2 + L_{\mathrm{cc}}\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}} = b_{\mathrm{cc}}\left(-DT(1-q)\right)^2 + \frac{1}{2}DT(1-q)^2DT \\ &= (DT)^2\left[b_{\mathrm{cc}}(1-q)^2 + \frac{1}{2}(1-q)^2\right] = (DT)^2(1-q)^2\left(b_{\mathrm{cc}} + \frac{1}{2}\right). \end{aligned}$$

It follows that $B2 \leq 0$ if $b_{\mathrm{cc}} \leq -0.5$.
Similarly for $B3$:

$$\begin{aligned} B3 &= b_{\mathrm{bc}}\left(\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2 + L_{\mathrm{cc}}\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}} = b_{\mathrm{bc}}\left(-DT(1-q)\right)^2 + \frac{1}{2}DT(1-q)^2DT \\ &= (DT)^2\left[b_{\mathrm{bc}}(1-q)^2 + \frac{1}{2}(1-q)^2\right] = (DT)^2(1-q)^2\left(b_{\mathrm{bc}} + \frac{1}{2}\right). \end{aligned}$$

It follows that $B3 \leq 0$ if $b_{\mathrm{bc}} \leq -0.5$.
Finally, for $B4$:

$$\begin{aligned} B4 &= b_{\mathrm{bo}}\left(\frac{\partial}{\partial q}L_{\mathrm{co}} + \rho\frac{\partial}{\partial q}L_{\mathrm{cc}}\right)^2 + \rho(L_{\mathrm{co}} + \rho L_{\mathrm{cc}})\frac{\partial^2}{\partial q^2}L_{\mathrm{cc}} \\ &= b_{\mathrm{bo}}(DT - \rho DT(1-q))^2 + \rho\left(qDT + \rho\frac{1}{2}DT(1-q)^2\right)DT \\ &= (DT)^2\left[b_{\mathrm{bo}}(1-\rho(1-q))^2 + \rho\left(q + \rho\frac{1}{2}(1-q)^2\right)\right]. \end{aligned}$$

It follows that $B4 \leq 0$ if $b_{\mathrm{bo}} \leq -\rho\frac{q+\rho\frac{1}{2}(1-q)^2}{(1-\rho(1-q))^2}$. It can be noticed that the behavior of $f(q) = -\rho\frac{q+\rho\frac{1}{2}(1-q)^2}{(1-\rho(1-q))^2}$ in the domain $[0,1]$ depends on ρ. If $\rho \leq 0.5$, then $f(q)$ is a non-increasing function and its minimum occurs when $q = 1$ ($f = -\rho$), i.e. $B4 \leq 0$ if $b_{\mathrm{bo}} \leq -\rho$. Moreover, when $\rho \leq 0.5$ and $-\rho < b_{\mathrm{bo}} < -\frac{1}{2}\frac{\rho^2}{(1-\rho)^2}$, then, for smaller values of q, $B4 < 0$, while for large values of q, $B4 > 0$. Inversely, if $0.5 < \rho < 1$, then $f(q)$ is an increasing function and its minimum occurs when $q = 0$ ($f = -\frac{1}{2}\frac{\rho^2}{(1-\rho)^2}$), i.e. $B4 \leq 0$ if $b_{\mathrm{bo}} \leq -\frac{1}{2}\frac{\rho^2}{(1-\rho)^2}$. Furthermore, when $0.5 < \rho < 1$ and $-\frac{1}{2}\frac{\rho^2}{(1-\rho)^2} < b_{\mathrm{bo}} < -\rho$, then, for smaller values of q, $B4 > 0$, while for larger values of q, $B4 < 0$.

Thus, when the steepness of the quantity discount is significant ($|b| > 0$), the terms $B2$, $B3$ and $B4$ in eq. (71) decrease, and consequently the sign of $\frac{\partial^2 C}{\partial q^2}$ changes

to negative, thereby resulting in a concave cost function. It is readily visible in a special case of a linearly distributed demand, $b_{\mathrm{co}} = b_{\mathrm{cc}} = b_{\mathrm{bo}} = b_{\mathrm{bc}} = b$ and $\rho \to 0$, for which the second derivative simplifies to:

$$\frac{\partial^2 C}{\partial q^2} = \frac{1}{(DT)^{b+1}(b+1)}\Big[bq^{b-1}(a_{\mathrm{co}} + a_{\mathrm{bo}}k^{b+1}) + (a_{\mathrm{cc}} + a_{\mathrm{bc}}k^{b+1})(1-q)^{2b}\left(\frac{1}{2}\right)^{b-1}\left(b+\frac{1}{2}\right)\Big] \tag{81}$$

As could be seen, in this special case, when $b \leq -\frac{1}{2}$, it follows that $\frac{\partial^2 C}{\partial q^2} < 0$.

As was illustrated above for the special case of the increased absolute values of b_{cc}, b_{bo}, b_{bc}, the open subsystem costs C may become a concave function of q. As a result, the minimum of costs is achieved with q at the edges of interval $[0, 1]$. In other words, when the unit price of the acquired capacity is subject to quantity discounting as reflected in the absolute values of b_{cc}, b_{bo}, b_{bc}, the minimum costs may be achieved by acquiring only private ($q = 1$) or only public capacity ($q = 0$), while the use of a hybrid cloud may be inefficient cost-wise, q.e.d. □

5 Illustrative Numerical Experiments

In the preceding section, the costs of a hybrid cloud-based service have been analytically explored. In particular, the effects of

- a non-constant demand for computing and data communication capacity,
- a varying intensity of data communication, and
- a quantity discount applied to the unit prices of computing and data communication capacity

were analyzed. In this section, some numerical examples, wherein these effects are modelled, are provided. These examples are aimed at illustrating how the above effects influence the costs of open subsystems, and in particular how they affect the cost-optimal distribution of acquired capacity among the private and the public clouds.

We now consider an imaginary case of a hybrid cloud-based service where the service provisioning to the customers requires both computational resources and some data communication overheads. The computing requirements are assumed to be fully satisfied by the equivalent of 500 Amazon EC2 large instances (Amazon Web Services, 2011), though this number may be changed without inflicting significant changes on the results of the experiments. Unless specified otherwise, a linear demand curve is assumed, i.e. the demand is uniformly distributed between zero and D as illustrated in Fig. 3.

For the linear demand curve (uniformly distributed demand), the cumulative acquired public cloud computing capacity is $L_{\mathrm{cc}} = \frac{1}{2}DT(1-q)^2$. The use of the linear demand curve, albeit unrealistic, allows finding the analytical solution to

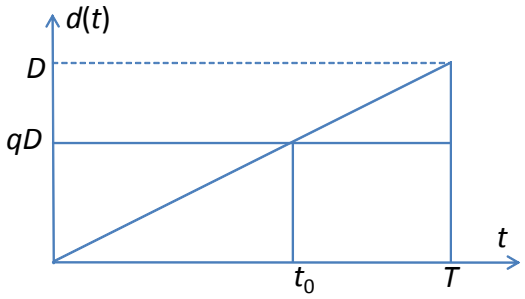

Fig. 3 Uniformly distributed demand curve rearranged to be monotonically non-decreasing

eq. (41) easily and thereby helps in illustrating some aspects of the proposed model. Meanwhile, since the Propositions 1 through 3 were shown to hold for an arbitrary demand distribution, the use of a more realistic demand distribution will not affect the results of the experiments illustrating those propositions. Proposition 5 relied on the special case of the linearly distributed demand function, and hence, the linear demand curve is employed also in the numerical experiments devoted to that proposition. Note that Proposition 4 focuses on the case of the constant demand function, and accordingly the constant demand function is assumed when illustrating that proposition.

It should be noted that data storage costs are not included as separate cost factors in the cost consideration. It is assumed that the storage provided, along with computing capacity, by a public cloud (for instance, Amazon EC2 large instance offers 850 GB of storage) is sufficient for service provisioning, whereas persistent storage, when needed, is provisioned in-house, as a part of the private cloud infrastructure.

The parameters are set to the following values:

- A 3-year period is considered, i.e. $T = 24 \times 365 \times 3 = 26280$ (hours).
- The computing demand is assumed to be fully satisfied with 500 Amazon EC2 large instances, i.e. $D = 500$.
- The volume of data transfer is measured in GB, i.e. $k = 1$ means that one working hour of a small EC2 instance requires 1GB of data to be transferred between the public cloud and the customers.

5.1 Constant prices of computing and communication capacity

Let us consider a case where the unit prices are not discounted and therefore can be seen as constant, i.e. $b_{\mathrm{co}} = b_{\mathrm{cc}} = b_{\mathrm{bo}} = b_{\mathrm{bc}} = 0$, as considered in Propositions 1 through 3. Using the pricing defined by Amazon for its EC2 services (Amazon Web Services, 2011), the prices of computing and communication capacities are set to the following values:

- The price of public cloud computing capacity is estimated based on the price of a standard, large on-demand Linux/UNIX instance located in EU: $p_{\mathrm{cc}} = 0.38$ (USD/hour).
- The price of public cloud data transfer is estimated based on the "Data Transfer Out" pricing of EC2 for US & EU Regions. If the total amount does not exceed 10TB/month, the price is fixed as: $p_{\mathrm{bc}} = 0.15$ (USD/GB).
- The price of public cloud capacity is provisionally assumed twice more expensive than that of the private cloud, i.e. $u_{\mathrm{c}} = u_{\mathrm{b}} = u = 2$ (Khajeh-Hosseini et al, 2011). Hence, $p_{\mathrm{co}} = 0.19$ and $p_{\mathrm{bo}} = 0.075$. Note that u can be changed without affecting the results, as long as $u > 1$.

The varying intensity of data communication is modelled by assigning different values to coefficients k and ρ: the larger the coefficient value, the greater the intensity.

Since quantity discounting is ignored, we can estimate the total cost according to eq. (26):

$$C = a_{\mathrm{co}} qDT + u a_{\mathrm{co}} L_{\mathrm{cc}} + a_{\mathrm{bo}} k(qDT + \rho L_{\mathrm{cc}}) + u a_{\mathrm{bo}} k(1+\rho) L_{\mathrm{cc}}. \tag{82}$$

Given that $L_{\mathrm{cc}} = \frac{1}{2} DT(1-q)^2$, we can rewrite the above as

$$\begin{aligned} C = {} & a_{\mathrm{co}} qDT + a_{\mathrm{bo}} k(qDT + \rho \frac{1}{2} DT(1-q)^2) \\ & + u \frac{1}{2} DT(1-q)^2 (a_{\mathrm{co}} + a_{\mathrm{bo}} k(1+\rho)). \end{aligned} \tag{83}$$

5.1.1 Negligible demand for data communication.

First, consider the case when the demand for communication capacity is low and can be ignored. In this case, $k = \rho = 0$, and the eq. (83) can be simplified to:

$$C = a_{\mathrm{co}} qDT + \frac{1}{2} u a_{\mathrm{co}} DT(1-q)^2. \tag{84}$$

In Fig. 4, the resulting costs of an open subsystem are plotted as a function of the threshold portion of private cloud demand q. As can be seen from the figure, in the absence of communication costs and quantity discounting, the minimum cost is achieved when a hybrid cloud is used, in line with Proposition 2. Furthermore, according to Corollary 1, the value of $q_{\min}$ minimizing the cost is determined by the ratio of the prices $q_{\min} = 1 - \frac{1}{u} = 1 - \frac{1}{2} = 0.5$.

5.1.2 Non-zero demand for data communication.

Let us now consider the effect of data communication on open subsystem costs. In Figure 5, the costs of an open subsystem are plotted as a function of q, given a set of different values of k and ρ. In the left part of the figure, the plots for different values of k are provided (the value of $\rho = 0.2$ is used). As can be seen,

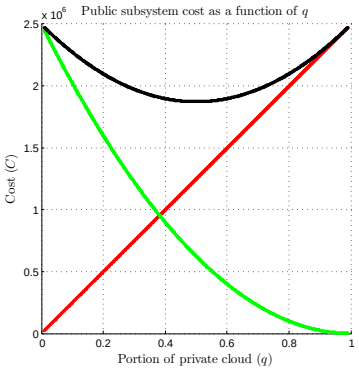

Fig. 4 The costs of private (red) and public (green) open subsystems, as well as the total cost of open subsystems (black). The unit prices of capacity are fixed, and the demand for communication capacity is negligible

the costs grow as the value of k increases (cf. Proposition 1). The value of $q_{\min}$, minimizing the costs (shown by vertical lines), shifts to the right as k increases, thus indicating that the greater the communication intensity, the more the private cloud capacity that should be acquired.

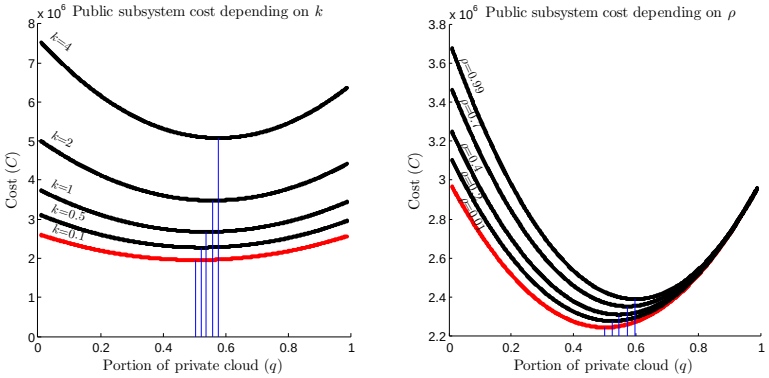

Fig. 5 The cost of open subsystems for different values of k (left) and ρ (right), plotted as a function of the private cloud demand threshold q. The vertical lines indicate the minimum costs for different values of k and ρ

The costs' dependency on the value of ρ depicted in the right part of the figure exhibits a similar pattern (the value of $k = 0.5$ is used). Namely, the costs grow with the value of ρ, and the value of $q_{\min}$, minimizing the costs, shifts to the right as ρ increases. Thus, the figure indicates that the greater the communication intensity between the organization's closed subsystems and the public cloud, the more the private cloud capacity that should be acquired, which agrees with Proposition 2.

Furthermore, it can be shown that:

$$\text{as } k \to +\infty,\ q_{\min} \to 1 - \frac{1}{u + \rho(1+u)} = 1 - \frac{1}{2 + 0.2 \times 3} = 0.6154 \tag{85}$$

$$\text{as } \rho \to +\infty,\ q_{\min} \to 1. \tag{86}$$

i.e. for larger values of ρ, the capacity should be mainly allocated to the private cloud.

Thus, for a linear demand curve, the data transfer between the organization and the public cloud has a greater impact on the cost-optimal distribution of acquired capacity than the communication between the open subsystems and the customers. In other words, it is more reasonable (cost-wise) to use the public cloud infrastructure for services which have little interaction with the closed in-house subsystems.

5.2 Non-constant prices of computing and communication capacity

In the preceding subsection, the costs were illustrated for the case of fixed prices of computing and communication capacity. In this subsection, the case of non-constant prices is considered by taking the quantity discount into account.

5.2.1 Pricing parameters

The pricing parameters are set to their values as follows:

For computing capacity

Public cloud. Amazon EC2 instances are priced equally, independently of how many instance-hours are consumed over the billing period. Also, when changing from a small standard to large or extra large instances, the charge per hour grows linearly with the number of EC2 computing units, i.e. no volume discounts are given. Therefore, the price of cloud computing capacity is assumed fixed, i.e. $b_{\text{cc}} = 0$. The price of a large EC2 instance is used for assigning the value of $a_{\text{cc}} = 0.38$ (USD/hour).

Private cloud. The in-house computing capacity acquisition costs are subject to the price elasticity of demand; Bayoumi and Haacker (Bayoumi and Haacker, 2002) measured the (absolute) price elasticity of demand for hardware to be in the range $|\varepsilon| = 1.1 \ldots 1.8$, and suggested that 1.3 is a "reasonable" value. However, assuming $b_{\text{co}} = 1/\varepsilon = -1/1.3 = -0.769$ would result in an incorrect estimate, since neither the underlying physical infrastructure nor the associated human costs are taken into account. Therefore, instead, parameters b_{co} and a_{co} are estimated as follows.

According to (Hamilton, 2010), when large ($N_{\text{CO1}} = 5 \times 10^4$ servers) and medium ($N_{\text{CO2}} = 10^3$ servers) datacenters are compared, the economies of scale

(Stigler, 1958) result in 5.7 . . . 7.1 difference in the network, storage, and administration costs. We take the value of $p_{\mathrm{CO2}}/p_{\mathrm{CO1}} = 6.5$ as a reasonable value. Since $p_{\mathrm{CO1}} = a_{\mathrm{co}}(N_{\mathrm{CO1}})^{b_{\mathrm{co}}}$ and $p_{\mathrm{CO2}} = a_{\mathrm{co}}(N_{\mathrm{CO2}})^{b_{\mathrm{co}}}$, it follows that $b_{\mathrm{co}} = \frac{\ln(p_{\mathrm{CO1}}/p_{\mathrm{CO2}})}{\ln(N_{\mathrm{CO1}}/N_{\mathrm{CO2}})}$. Therefore, $b_{\mathrm{co}} = \frac{\ln(1/6.5)}{\ln(50000/1000)} = -0.478$.

The estimates by Greenberg et al. (Greenberg et al, 2008)[2] suggest that the cost of a large datacenter with 5×10^4 servers is 5 621 117 (USD/month). Assuming 50 virtual machines (VM) per server, such datacenter may host $N_{\mathrm{VM}} = N_{\mathrm{co}} \times 50 = 2.5 \times 10^6$ VMs. Then, the cost of one VM per hour is $p_{\mathrm{co}} = 5\ 621\ 117/(2.5 \times 10^6 \times 30 \times 24) = 3.1 \times 10^{-3}$. Since $p_{\mathrm{co}} = a_{\mathrm{co}}(N_{\mathrm{VMH}})^{b_{\mathrm{co}}}$, where $N_{\mathrm{VMH}} = N_{\mathrm{VM}} \times T$ is the number of VM-hours provided by the datacenter over time T, it follows that $a_{\mathrm{co}} = p_{\mathrm{co}}(N_{\mathrm{VMH}})^{-b_{\mathrm{co}}} = 415.93$.

For data communication capacity

Public cloud. Based on the pricing of Amazon EC2 (Amazon Web Services, 2011), the parameters a_{bc} and b_{bc} are estimated by using the least-square fitting as: $a_{\mathrm{bc}} = 0.773$ and $b_{\mathrm{bc}} = -0.130$.[3]

Private cloud. The price of internet connection in-house is usually set using one of the following three methods (Stiller et al, 2001): i) a fixed monthly charge depending on the allocated bandwidth, ii) a volume-based charge, or iii) a bursty rate depending on the 95% highest sample of consumed bandwidth. The first method is arguably the most widely used by ISPs (Stiller et al, 2001; Odlyzko, 2001), therefore, it is assumed.

Furthermore, monthly fees are assumed to grow non-linearly with the allocated bandwidth (Opitz et al, 2008), and hence the unit price of reserved data communication capacity is assumed to be subject to a quantity discount. For simplicity, the unit price (per GB) is approximated with the same parameters as were obtained for the Amazon data transfer prices, i.e. $a_{\mathrm{bo}} = a_{\mathrm{bc}} = 0.773$ and $b_{\mathrm{bo}} = b_{\mathrm{bc}} = -0.130$[4].

Assuming the pricing with the parameters described above, the cost are considered below i) for the constant demand function and ii) for the uniformly distributed demand function.

[2] Available at `http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx`

[3] Note that, due to the pricing scheme of Amazon, this fitting was done for the large data communication volumes (exceeding 10TB per month). Therefore, these parameters give somewhat incorrect results for the volumes less than 10TB per month.

[4] In fact, the data-communication price for an enterprise would be determined by the overall communication capacity used in the enterprise: $L_{\mathrm{bo}} + L_{\mathrm{boc}} + L_0$, where L_0 is the data communication capacity used by all other services in the enterprise. Here, for simplicity it is assumed that L_0 is small compared with $L_{\mathrm{bo}} + L_{\mathrm{boc}}$ and hence can be ignored.

5.2.2 Constant demand.

First, let us consider the case of constant demand function $d(t) = D$. Then, $L_{\text{co}} = qDT$ and $L_{\text{cc}} = (1-q)DT$, and therefore, the open subsystem cost

$$C = a_{\text{co}} L_{\text{co}}^{b_{\text{co}}+1} + a_{\text{cc}} L_{\text{cc}}^{b_{\text{cc}}+1} + a_{\text{bo}}[k(L_{\text{co}} + \rho L_{\text{cc}})]^{b_{\text{bo}}+1} + a_{\text{bc}}[kL_{\text{cc}}(1+\rho)]^{b_{\text{bc}}+1}$$

can be rewritten as

$$C = a_{\text{co}}(qDT)^{b_{\text{co}}+1} + a_{\text{cc}}((1-q)DT)^{b_{\text{cc}}+1} + a_{\text{bo}}[k(qDT + \rho(1-q)DT)]^{b_{\text{bo}}+1} + a_{\text{bc}}[k(1-q)DT(1+\rho)]^{b_{\text{bc}}+1} \quad (87)$$

The costs of open subsystems for this case are depicted in Fig. 6. As can be seen, the graphs of the cost function are concave. This is in line with our reasoning in the previous section, where it was shown (cf. Proposition 4) that, given a flat demand distribution function, the cost of a private and/or public cloud infrastructure are lower as compared with a hybrid cloud.

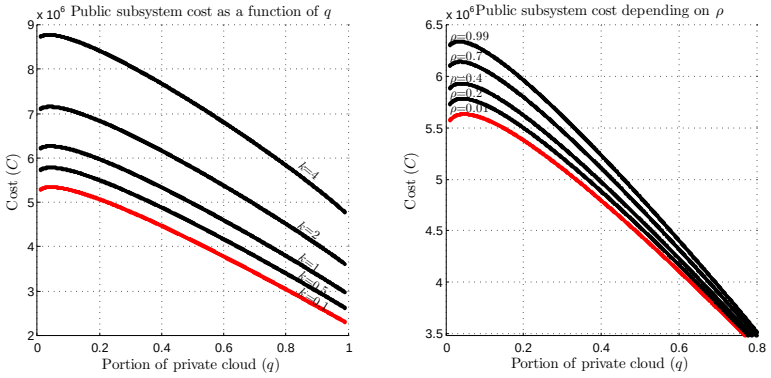

Fig. 6 The cost of open subsystems for different values of k (left) and ρ (right), plotted as a function of the private cloud demand threshold q. A constant demand and a non-zero steepness of the quantity discounting are assumed

Furthermore, the figure also illustrates that, in line with Corollaries 3 and 4, while the costs increase with communication overhead (k and ρ), the minimum still occurs at $q = 1$, i.e. when the in-house infrastructure is used. Thus, the minimum costs are achieved when the private cloud only is used.

5.2.3 Linear demand distribution function (uniformly distributed demand).

Let us now consider the case of the linear demand distribution curve, i.e. the case when $L_{\text{cc}} = \frac{1}{2}DT(1-q)^2$.

In Fig. 7, the costs of computing and the costs of data communication are shown separately. The left plot in the figure illustrates the computing capacity

cost as a function of q. As can be seen, the computing capacity cost function is neither convex nor concave; rather, the function is concave in the area of small q values and convex for the remaining values of q. This is due to the mutual effect of non-constant demand (convex cost function constituent, cf. Proposition 2), and quantity discounting (concave cost function constituent considered in Propositions 4 and 5).

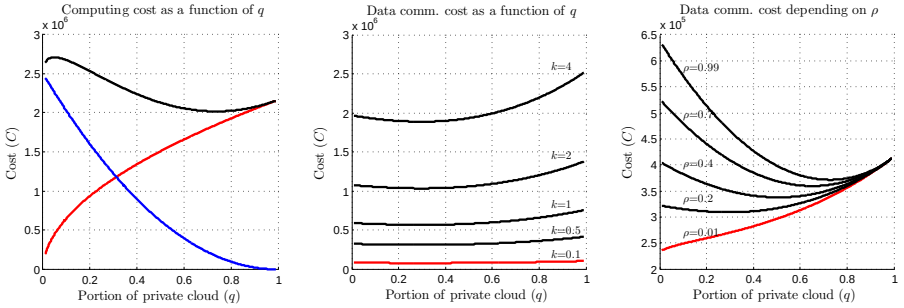

Fig. 7 The cost of open subsystems plotted as a function of the private cloud demand threshold q. The cost of computing capacity is shown in the left, where the costs of the private (red), public (blue), and total (black) computing capacity are plotted. The middle and the right figures portray the data communication costs for different values of k (middle) and ρ (right). A linearly distributed demand and a non-zero steepness of the quantity discount are assumed.

Data communication capacity cost (shown in the middle and in the right) is largely a convex function, as was discussed in Section 4. However, as can be seen from the figure, the data communication costs, too, are affected by quantity discounting (concave cost function constituent): for instance, for $\rho = 0.01$, the cost function is concave in the area of small q values.

According to Fig. 7, the concavity is manifested largely in the computing capacity cost, whereas it is almost non-present in the data-communication costs. The difference in concavity suggests that quantity discounting (concave cost function constituent) affects more the computing capacity costs than the data communication capacity costs – this is due to the fact that the steepness of the quantity discount for data communication capacity is lower (in absolute value), as compared with the steepness for the computing capacity, and hence contributes less to the cost function.

In Fig. 8, the resulting costs of open subsystems are shown. As the plots in the figure illustrate, due to the mutual effect of non-constant demand (convex cost function component), and quantity discounting (concave cost function component), the resulting cost functions are neither concave nor convex. Rather, they are concave in the area of small q values and convex for the remaining values of q. This change from concavity to convexity (as the q values increase) indicates that

the second derivative $\frac{\partial^2 C}{\partial q^2}$ changes its sign from negative to positive, as the term $B4$ in eq. (71) grows (cf. the proof of Proposition 5).

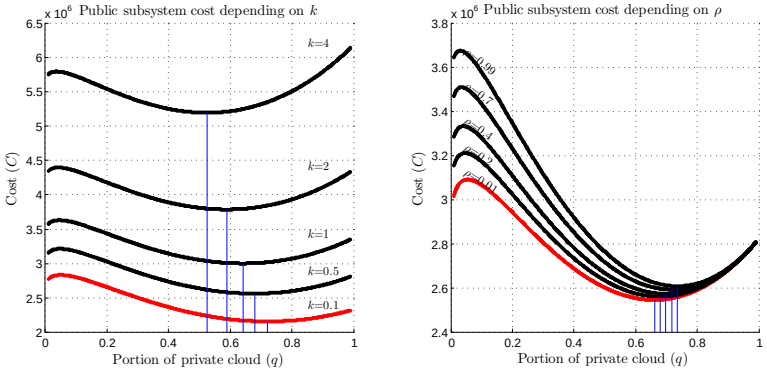

Fig. 8 The cost of open subsystems for different values of k (left) and ρ (right), plotted as a function of the private cloud demand threshold q. The vertical lines indicate the minimum costs for different values of k and ρ. A linearly distributed demand and a non-zero steepness of the quantity discount are assumed

As the figure indicates, a hybrid cloud is cost-optimal ($0 < q_{\min} < 1$). The values of $q_{\min}$ are found within the areas where the cost function is convex, thus indicating that the volume of non-constant demand has a decisive effect on the overall costs, outbalancing the effect of the quantity discounting.

It can be observed, that, as the data communication intensity grows (i.e. k and ρ increase), the effect of the non-constant demand for data communication (the convex cost function component) starts to dominate – and as a result, the areas, wherein the cost function is convex, enlarge. As discussed above, this can be explained by the fact that the steepness of the quantity discount for data communication capacity is lower, in absolute value, as compared with the steepness for the computing capacity; therefore, as the portion of the data-communication costs in the overall costs increases, the portion of the computing capacity costs drops, and hence the effect of data communication costs – mainly convex – starts to dominate.

It is noteworthy that, as k increases, the value of $q_{\min}$ shown with blue vertical lines in the figure decreases (whereas $q_{\min}$ increases with ρ). This is in line with Corollary 2 stating that in case $u_c \gg u_b$, $q_{\min}$ decreases as k grows. Indeed, if, for example, 250 instances are deployed in the private cloud, then $p_{co} = 415.93 \times (250 \times 24 \times 365 \times 3)^{-0.478} = 0.23$, and hence $u_c = \frac{p_{cc}}{p_{co}} = \frac{0.38}{0.23} = 1.65$, while $u_b = 1$. As a result, the term $\rho(1 + u_b) + u_b - u_c = 0.2 \times (1 + 1) + 1 - 1.65 = -0.25$ in the partial derivative $\frac{\partial \tau_c}{\partial k}$ is negative, thus resulting in $\frac{\partial Q}{\partial k} < 0$, and hence $q_{\min}$ decreases as k grows.

Above, the cost function curve was shaped by a mutual effect of the non-constant demand and the quantity discounting. In order to study the effect of the absolute value of the steepness of the quantity discount on the total costs, the total costs are shown for a set of values of the steepness in Fig. 9:

$$b = \{0.0; -0.075; -0.125; -0.25; -0.5; -0.6\}.$$

In the list, the value of $b = -0.125$ approximates the steepness of the quantity discount for data communication (-0.130), whereas the value of $b = -0.5$ approximates the quantity discounting of the computing capacity (-0.478). The data communication parameters are set to values $k = 0.5$ and $\rho = 0.2$. Furthermore, the public cloud discounting parameters are set to the same values as in the private cloud (to avoid the zero-discounting computing capacity prices as set by Amazon).

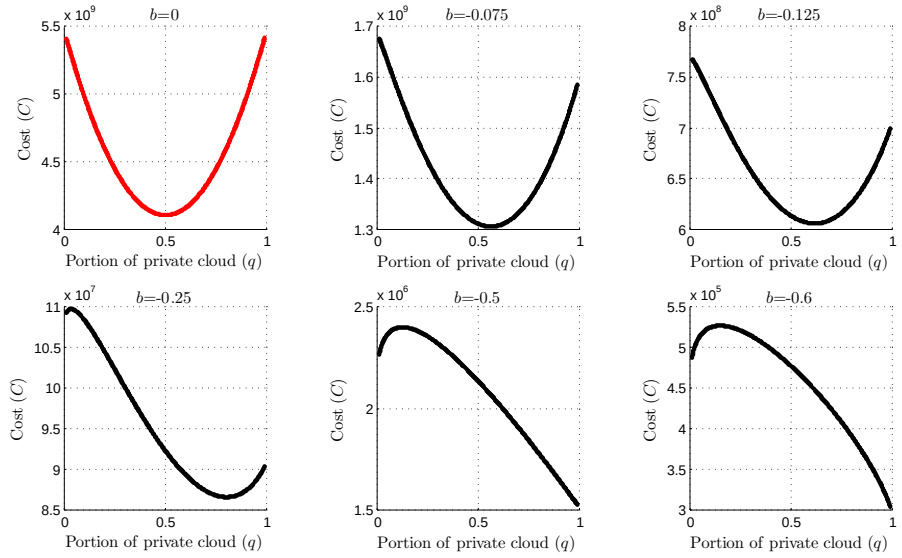

Fig. 9 The cost of open subsystems for different values of the steepness of the quantity discount ($b = \{0.0; -0.075; -0.125; -0.25; -0.5; -0.6\}$), plotted as a function of the private cloud demand threshold q. A linearly distributed demand is assumed

As can be seen, the cost function is convex when the steepness of the quantity discount is low. However, the greater the absolute value of the steepness, the less the convexity. Eventually, as the absolute value of the steepness increases, the cost function becomes partially concave (visible already for $b = -0.125$), and then fully concave ($b = -0.5$); in other words, the cost function concavity grows with the absolute value of the steepness of the quantity discount. As discussed above, this explains the more concave shape of the cost function when the data communication intensity is low, and more convex shape when the data communication intensity increases.

The pricing schemes present today exhibit different discounting for the computing capacity and for the data communication capacity. Fig. 9 manifests possible future scenarios when, due to technological advances and market trends, approximately the same quantity discount applies to the both types of capacity. On the one hand, if the quantity discount is at the low level as observed currently in the data communication capacity pricing ($b = -0.125$), the price function will likely be convex with the minimum at $q \in (0, 1)$ suggesting the use of the hybrid cloud infrastructure. On the other hand, should the steepness of the quantity discount for data communication capacity reach the level of $b = -0.5$ similarly to the quantity discounting of today's private computing capacity, the cost function will likely be concave, with the minimum achieved at $q = 1$, suggesting the use of the private cloud only.

6 Discussion

In previous sections, the model for hybrid cloud costs was introduced. In it, the costs of computing capacity and data communication capacity are modelled as a function of the threshold demand for computing capacity. Whereas the demand up to this threshold value is served with the private cloud infrastructure, the demand exceeding the threshold value is served with the public cloud infrastructure. This model can be employed for identifying the cost-optimal division between the private and the public capacity, as was illustrated with the help of numerical experiments. Below, some theoretical and practical implications of the proposed model are considered, and the directions for further work are outlined.

6.1 Theoretical and practical implications

The findings reported in this paper have some implications on the research of plural governance and concurrent sourcing:

Diseconomies of scale due to *volume uncertainty* combined with the costs of unused excess capacity have been considered as one of the hypothetical reasons for concurrent sourcing (Puranam et al, 2006). The effect of volume uncertainty has been taken into account in this paper by considering the form of demand function. For instance, it was shown that, given a non-constant demand function, the time of using the public cloud capacity should be the inverse of the premium charged by the cloud infrastructure vendor. Thus, the results of the paper provide analytical evidence supporting the above hypothesis in the context of the concurrent sourcing of computing infrastructure.

Economies of scale – i.e. the reduction of the average cost per unit of a good/service with the number of units produced (Stigler, 1958) – have been referred to as a factor "increasing the likelihood that the production is kept internally", thus suggesting the use of single sourcing (Mols, 2010). In the context of cloud infrastructure,

the economies of scale are manifested in the quantity discount applied to the unit price of an infrastructure capacity. It has been shown in the paper that, given a constant demand, the effect of quantity discounting – and hence the economies of scale – renders the use of hybrid cloud unreasonable. Therefore, the analysis in the paper is in line with the above claim that the economies of scale make single sourcing the preferred option.

While the *transaction cost theory* does not address the phenomena of concurrent sourcing (Mols, 2010), it does consider the use of internal vs. external production alternatives and suggests that transaction-specific costs make the internal mode of production more likely. This paper exemplifies that the concurrent sourcing may incur *extra costs*, in this case the costs of data communication between the in-house and cloud subsystems (L_{boc}), which are needed for coordination and interaction between in-house and outsourced parts. These extra costs make the concurrent sourcing threshold high whenever the outsourced process and the in-house processes are tightly coupled. In the context of computing infrastructure, the extra costs of data communication play the role similar to the *transaction-specific costs* in the transaction cost theory: these costs are only incurred due to the collaboration within the hybrid cloud infrastructure, and therefore are highly transaction-specific, thus making the in-house mode of infrastructure more likely.

Importantly, whereas the transaction-specificity is often difficult to quantify, and hence elaborate questionnaire tools are usually employed in order to estimate it, the extra data communication expenses manifesting the transaction-specific costs in the context of computing infrastructure can be quantified directly, thus providing a unique opportunity to study the effect of the transaction-specific costs on sourcing decisions.

From the perspective of a practitioner, the proposed model enables the analysis and identification of a cost-efficient allocation of computing and data communication capacity to the in-house and public infrastructure, depending on both the form of the demand curve and the available pricing for the computing and the data communication capacities. As was illustrated numerically for a specific case of a communication-intensive application, a hybrid cloud may have up to $10-30\%$ lower costs than a fully private or a fully public cloud solution (cf. Fig. 8). Meanwhile, in case of intensive communication and a constant demand, the use of a hybrid cloud can have up to $40-60\%$ higher costs as compared with the in-house operation (cf. Fig. 6).

Noteworthy, the identification of the cost-efficient allocation requires a rather modest set of computations to be performed, using the historical information about capacity demand distribution and the information on pricing as an input. Furthermore, such estimation needs to be performed infrequently (e.g. when pricing parameters change dramatically), and hence the computational overheads of the estimation process are negligible.

The model also provides a possibility to explore possible future scenarios when, due to technological advances and market developments, the steepness of the quantity discount for a capacity changes. On the one hand, if the steepness becomes low, as observed currently for the data communication capacity, the cost function will likely be convex with the minimum at $q \in (0, 1)$ suggesting the use of the hybrid cloud infrastructure. On the other hand, should the steepness of the quantity discount for the data communication capacity reach the level of $b = -0.5$ – which would be similar to the steepness of today's private computing capacity – the cost function will likely be concave, with the minimum achieved at $q = 1$, thus suggesting the use of the private cloud only.

6.2 Limitations and further research

The analysis in this paper has focused on the cost of computing capacity and data communication costs, both of which depend on the size of the portion of demand that is served by the private/public cloud. There is a difference between the effects of these two factors on the overall costs. Namely, when the demand is moved from the private to the public subsystems (i.e. when q decreases):

- The cost of private cloud computing capacity decreases linearly, and the cost of public cloud computing capacity increases proportionally to the time when the public capacity is used.
- The cost of private cloud data communication also decreases; however, the decline is not linear due to the need for communication between the private and public cloud subsystems (reflected in the value of ρ).

For many cloud applications, such as online image processing systems, the two cost factors above constitute the majority of their computing infrastructure costs. In other application scenarios, depending on the system architecture and functionality provided, also other factors, such as the cost of the load balancer and the cost of persistent data storage, may contribute to the overall costs of the hybrid cloud. The contribution of additional cost factors to the overall costs depends on whether these factors are attributable to a single subsystem (either private or public, as computing capacity costs) or to an interaction between the private and public clouds (as data communication costs).

For instance, in systems with an excessive demand for persistent storage the effect of the storage cost depends on the replication of storage between the private and public clouds. In case the replication is not needed, the storage cost is expected to have an effect similar to that of the computing costs, i.e. it is expected to decrease linearly with q in the private cloud and increase proportionally to the time of use in the public cloud. However, if the public cloud's storage is replicated in the private cloud, the storage cost are expected to behave similarly to the data communication cost. The analysis of the storage costs is further complicated by the

fact that the pattern of using the persistent storage capacity is likely to differ from the computing or data communication capacities: whereas the computing or data communication resources are released once the computing task or communication is completed, the data may need to be stored for a long period of time resulting in incrementally increasing demand for the persistent storage capacity. Thus, the cost of persistent storage is a complex function determined by multiple factors, and therefore further research aimed at clarifying the contribution of storage expenses to the overall costs is worthwhile.

Some of the additional cost factors, such as the cost of a load balancing element, can be assumed independent of the specific load division between the private and the public clouds, and hence may be ignored when seeking a cost-efficient division. Still, the load balancing algorithm affects the pattern of allocating and releasing public cloud resources and hence influences the public cloud costs (den Bossche et al, 2010; Genaud and Gossa, 2011). Therefore, the details of the applicable load distribution algorithms, their effect on the public cloud costs, and associated computational overhead shall be studied as a part of future work.

Another aspect that warrants consideration in further research is the process of transforming legacy system architecture so that the hybrid cloud deployment would be enabled. Such a transformation may require additional system elements to be implemented, deployed and integrated, bringing additional costs and constraints. As a result, the cost advantage of adopting a hybrid cloud solution may decrease.

Finally, further research shall be devoted to the elaboration of a general cloud cost framework, wherein various cost factors would be categorized according to their contribution to the overall costs. For instance, the costs can be categorized into i) fixed costs, such as the costs of closed subsystems, ii) the costs incurred by either the private or the public portion of the cloud, such as the computing capacity costs, and iii) the costs incurred due to the interaction of the private and the public clouds, as exemplified by the communication costs. When integrating these costs, the framework shall also take into account the relative importance of individual factors, which depend on the form of the demand distribution for a specific resource. The aspects, such as trends in pricing and the net present value (NPV) of money, could also be taken into account in this framework.

7 Conclusions

The use of cloud infrastructure promises enterprises a reduction in IT costs, as well as faster time to market and improved scalability. Among different cloud infrastructure deployment modes, the hybrid mode is often argued to be more cost-efficient than either the private or the public cloud, due to the possibility of supplementing the limited capacity of private infrastructure with the capacity of the public cloud, when needed. In order to minimize the costs of such hybrid

cloud, a balance between the reserved private cloud capacity and acquired public cloud capacity should be found; in other words, the higher price of the public cloud capacity should be balanced with the relatively short duration of the time when the public cloud is utilized.

In this paper, a model for hybrid cloud costs, encompassing the costs of computing capacity and data communication capacity, has been introduced. In the proposed model, the costs are modelled as a function of the threshold demand for computing capacity, which is provided with the private cloud. The demand up to this threshold value is served with the private cloud infrastructure, which is assumed to be acquired beforehand and reserved for the purposes of service provisioning; whenever the demand exceeds the threshold value, the exceeding portion of the demand is served with the public cloud infrastructure, which is used without a prior reservation (on-demand) and charged based on the actual usage. When estimating the costs of a capacity, quantity discounting is taken into account. Using the model, the cost-optimal threshold for dividing the private and the public cloud computing capacity can be identified. Finding such optimal division has been numerically exemplified for the case of a demand uniformly distributed between zero and maximum levels.

It has been analytically shown that when the unit prices are fixed:

- A hybrid cloud may have lower costs than a fully private cloud or a fully public cloud solution;
- The presence of data communication costs shifts the cost-optimal division towards the private cloud, i.e. the greater the communication intensity, the more the private cloud capacity that should be acquired; and
- In the absence of data communication overheads, and given an arbitrary monotonically increasing demand distribution function, the portion of the time when public cloud is used should be the inverse of the premium charged by the cloud infrastructure vendor.

On the other hand, when the unit prices are subject to quantity discounting (i.e. decrease with the amount of acquired capacity),

- A non-hybrid solution – i.e. private or public cloud infrastructure, but not a hybrid solution – may provide the minimal costs.
- Given a constant demand, a fully in-house deployment provides the minimum costs.

A series of numerical experiments were employed in order to illustrate the above effects. In these experiments, the cost of open subsystems was plotted as a function of q – the threshold demand provided with the private cloud infrastructure.

The numerical experiments supported the claim that, under the condition of zero quantity discount, hybrid cloud minimizes the overall costs of the open subsystems. It was also evidenced by the experiments that the data transfer – either between the organization and the public cloud or between the private/public cloud

and the customers – increases the cost-optimal threshold for computing capacity to be provided with the private cloud. Also, the data transfer between the organization and the public cloud was found to have a greater impact on the cost-optimal distribution of acquired capacity in case of uniformly distributed demand. From practitioners' viewpoint this suggests that the services provided from the public cloud should avoid excessive communication with back-office systems.

The experiments also emphasized the effect of the quantity discount on the overall costs of open subsystems. As was shown in the case where the demand for computing capacity was distributed uniformly, due to the quantity discounting of computing and data-communication capacities, the overall cost may become a concave function of the private cloud threshold. As a result, the use of a hybrid cloud becomes economically unreasonable, since the cost is minimized by using a private or a public cloud alone.

In summary, the introduced model contributes to the previous work in the domain of the economics of cloud computing by taking into account the data communication overheads when estimating the costs of a hybrid cloud, and by taking into account quantity discounting. In future work, this model could be expanded towards a general cloud cost framework, where the other cost factors, such as the costs of public cloud data storage and the control cost incurred during the process of introducing hybrid cloud into the organization, would be taken into account.

Acknowledgements The research reported in this paper was carried out within the framework of the Cloud Software Program which was governed by TIVIT Oy nominated to organize and manage the programs of the Strategic Center for Science, Technology and Innovation in the field of ICT funded by the Finnish Funding Agency for Technology and Innovation (TEKES).

References

Amazon Web Services (2011) Amazon elastic compute cloud (amazon ec2). Available from `http://aws.amazon.com/ec2/`, retrived on 15.5.2011

Bayoumi T, Haacker M (2002) It's not what you make, it's how you use it: Measuring the welfare benefits of the it revolution across countries. Discussion paper, Centre for Economic Performance, London School of Economics and Political Science

den Bossche RV, Vanmechelen K, Broeckhove J (2010) Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads. In: 3th IEEE International Conference on Cloud Computing, IEEE Computer Society, Los Alamitos, CA, USA, pp 228–235, DOI http://doi.ieeecomputersociety.org/10.1109/CLOUD.2010.58

Genaud S, Gossa J (2011) Cost-wait trade-offs in client-side resource provisioning with elastic clouds. In: IEEE (ed) 4th IEEE International Conference on Cloud Computing (CLOUD 2011)

Greenberg A, Hamilton J, Maltz DA, Patel P (2008) The cost of a cloud: research problems in data center networks. SIGCOMM Comput Commun Rev 39:68–73, DOI http://doi.acm.org/10.1145/1496091.1496103, URL http://doi.acm.org/10.1145/1496091.1496103

Hamilton J (2010) Cloud computing economies of scale. Keynote at AWS Genomics & Cloud Computing Workshop, Seattle, WA, 08.06.2010, available from http://www.mvdirona.com/jrh/TalksAndPapers/JamesHamilton_GenomicsCloud20100608.pdf

Harms R, Yamartino M (2010) The economics of the cloud. Whitepaper, Microsoft

Heide JB (2003) Plural governance in industrial purchasing. Journal of Marketing 67:18–29

Jung JJ (2010) Reusing ontology mappings for query segmentation and routing. Semantic Peer-to-Peer Environment, Information Sciences 180(17):3248–3257

Khajeh-Hosseini A, Greenwood D, Smith JW, Sommerville I (2011) The cloud adoption toolkit: supporting cloud adoption decisions in the enterprise. Software: Practice and Experience - Special Issue on Software Architectures and Application Development Environments for Cloud Computing DOI 10.1002/spe.1072, URL http://dx.doi.org/10.1002/spe.1072

Kotler P, Keller K (2008) Marketing Management. Prentice Hall

Lee CA (2010) A perspective on scientific cloud computing. In: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, ACM, New York, NY, USA, HPDC '10, pp 451–459, DOI http://doi.acm.org/10.1145/1851476.1851542, URL http://doi.acm.org/10.1145/1851476.1851542

Mazhelis O, Tyrvinen P (2011) Role of data communications in hybrid cloud costs. In: Proceedings of the 37th EUROMICRO Conference on Software Engineering and Advanced Applications

Mell P, Grance T (2010) The nist definition of cloud computing. Version 15, 10-7-09, National Institute of Standards and Technology, available from http://www.csrc.nist.gov/groups/SNS/cloud-computing/

Mols NP (2010) Economic explanations for concurrent sourcing. Journal of Purchasing and Supply Management 16(1):61 – 69, DOI DOI:10.1016/j.pursup.2009.09.001, URL http://www.sciencedirect.com/science/article/pii/S1478409209000624

Odlyzko AM (2001) Internet pricing and the history of communications. Computer Networks 36(5-6):493–517

Opitz A, König H, Szamlewska S (2008) What does grid computing cost? J Grid Comput 6(4):385–397

Parmigiani A (2007) Why do firms both make and buy? an investigation of concurrent sourcing. Strategic Management Journal 28(3):285–311, DOI 10.1002/smj.580, URL http://dx.doi.org/10.1002/smj.580

Perloff J (2008) Microeconomics: Theory and Applications with Calculus. Prentice Hall

Porter M (1980) Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, New York

Pring B, Brown R, Frank A, Hayward S, Leong L (2009) Forecast: Sizing the cloud; understanding the opportunities in cloud services. Tech. rep., Gartner Dataquest

Puranam P, Gulati R, Bhattacharya S (2006) How much to make and how much to buy: An analysis of optimal plural sourcing strategies, working paper, available at SSRN: http://ssrn.com/abstract=932606 (last retrieved on March 10, 2011)

Risch M, Altmann J (2008) Cost analysis of current grids and its implications for future grid markets. In: Altmann J, Neumann D, T F (eds) Proceedings of the 5th international workshop on Grid Economics and Business Models, Springer-Verlag, Berlin, Heidelberg, GECON '08, pp 13–27, DOI http://dx.doi.org/10.1007/978-3-540-85485-2_2, URL http://dx.doi.org/10.1007/978-3-540-85485-2_2

Schotanus F, Telgen J, de Boer L (2009) Unraveling quantity discounts. Omega 37(3):510 – 521, DOI DOI:10.1016/j.omega.2007.09.002, URL http://www.sciencedirect.com/science/article/pii/S0305048307001004

Shah NH, Dixit VM (2005) Price discount strategies: a review. Revista investigacion operacional 26(1)

Stigler G (1958) The economies of scale. Journal of Law and Economics 1:54–71

Stiller B, Reichl P, Leinen S (2001) Pricing and cost recovery for internet services: Practical review, classification and application of relevant models. NETNOMICS - Economic Research and Electronic Networking 3(1)

Stole LA (2003) Handbook of industrial organization, North-Holland, chap Price Discrimination and Imperfect Competition

Strebel J, Stage A (2010) An economic decision model for business software application deployment on hybrid cloud environments. In: Schumann M, Kolbe LM, Breitner MH, Frerichs A (eds) Multikonferenz Wirtschaftsinformatik 2010, Universittsverlag Gttingen, p 195206

Weinman J (2011) Mathematical proof of the inevitability of cloud computing, working paper, available from http://www.joeweinman.com/Resources/Joe_Weinman_Inevitability_Of_Cloud.pdf (last retrieved on March 10, 2011)

Williamson O (1985) The Economic Institutions of Capitalism. The Free Press, New York

Youseff L, Butrico M, Da Silva D (2008) Toward a unified ontology of cloud computing. In: 2008 Grid Computing Environments Workshop ((GCE'08)), IEEE, pp 1–10, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4738443