Role of Data Communications in Hybrid Cloud Costs

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

Abstract—**Rapid adoption of cloud services in recent years has been driven by multiple factors, such as faster time-to-market and improved scalability enabled by public cloud infrastructure. Hybrid clouds, combining the in-house capacities with on-demand capacity of public clouds, achieve both the increased utilization rate of the in-house infrastructure and the limited use of more expensive public cloud, thereby lowering the total costs for the cloud user. 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, the 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 is shown analytically that the greater the volume of data transferred to/from the public cloud, the greater portion of the capacity should be allocated to the private cloud.**

Keywords-hybrid cloud; cost model; data communications cost

I. INTRODUCTION

Cloud computing represents a state-of-the-art "computing as a service" paradigm, where the configurable computing resources are pooled and shared among multiple users and are efficiently provisioned to them on-demand through a broadband network access [1]. The deployment of cloud services promises enterprises a number of benefits, such as faster time to market and improved scalability [2], as well as cost benefits in terms of lower start-up and/or operations costs [3][4]. Owing to these promises, the cloud services are adopted rapidly: according to Gartner, the market for cloud services exceeded $46 billion in 2009 and will reach $150 billion by 2013 [5].

Four modes of cloud deployment can be recognized [1]: private, community, public, and hybrid. A private cloud is operated by a single organization itself, 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, which are able to avoid large up-frond IT investments. A hybrid cloud represents a combination of a public cloud with the organization's private cloud and is aimed at efficient distribution of the load among the clouds.

By supplementing the local infrastructure with computing capacity from a public cloud, the hybrid cloud enables organizations to increase the utilization of their reduced IT infrastructure and thereby reduce their IT costs. As argued by Weinman [3], a hybrid cloud is often more cost-efficient than the private cloud, since the high premium charged by the public cloud provider is compensated by the relatively short duration of the load peaks when the public cloud is utilized. Furthermore, it is shown in [3] that 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. The cost-optimal load distribution found in [3] assumes that only the computing capacity is charged for by the cloud service provider, and 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.

The claim on cost-efficiency of the hybrid cloud is partially confirmed in [6], where the conclusion made is that the usage of a computing grid infrastructure is economically advantageous when the demand for computing exhibit infrequent (once every several month) peaks that can be covered with grid capacities. Ref. [7] focuses on the cost-efficient mix of internal and external computing resources in a hybrid cloud; in the proposed approach, individual applications are assigned to either internal or external resources using mixed-integer programming. Based on the simulation results, the authors find that off-loading peak demand to the cloud may not bring any cost-benefits to the clients, though the authors acknowledge the preliminary nature of the findings and suggest the need for further research in this direction.

The phenomena related to the concurrent use of in-house and external capacity – namely, tapered integration [8], plural governance [9], concurrent sourcing [10] – has been studied in the strategic management literature; see [11] for a comprehensive review. It was found that in the markets characterized by demand uncertainty, the risk of diseconomies of scale due to unutilized excessive capacity may be mitigated by scaling down internal capacity, and supplementing it during peak demand with the externally acquired capacity [9] [12]. However, the efficient concurrent use of in-house and on-demand computing capacity was not addressed in these studies.

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 is introduced in the paper, wherein the costs of computing and data communication are taken into account. Using the model, the cost-optimal load division can be identified, as exemplified for the case of demand uniformly distributed between zero and maximum. It is shown analytically that, given an arbitrary demand distribution, 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 portion of the demand should be allocated to the private cloud.

A final draft version of the paper: Mazhelis, O. and Tyrväinen, P., "Role of Data Communications in Hybrid Cloud Costs," in Biffl, s., Koivuluoma, M., Abrahamsson, P., and Oivo, M. (Eds.) SEAA2011, Proc. of the 37th EUROMICRO Conference on Software Engineering and Advanced Applications, IEEE Computer Society, 2011. pp.138-145.

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 assumptions made are listed. The analytical model is introduced in Section 3. Numerical experiments illustrating the effect of data communication costs are described in Section 4. Finally, conclusions to the paper are given in Section 5.

II. Hybrid Cloud

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 of *open subsystem* is employed in order to emphasize the fact that the subsystems' 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 for the observable future.

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 substantial volume of data to be transferred, as depicted in the figure by using bold arrows.

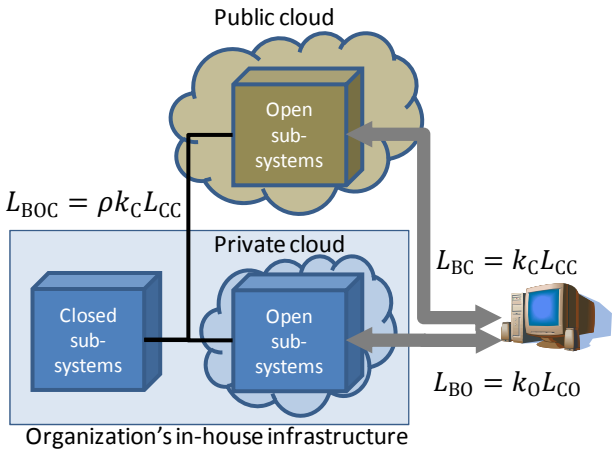

Figure 1. System decomposition according to the subsystem deployment forms

The proposed model is focused on the costs related to the open subsystems, with the purpose to identify the distribution of open subsystems between the private and the public cloud which would minimize the costs. The costs of closed subsystems do not depend on 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. Two cost components are considered:

- The costs of computing capacity, such as hardware, software, and data storage; and
- Data communications 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 under-utilized resources.

Similarly to [3], the following assumptions are made:

- Public cloud capacity is paid for only when used;
- The unit cost of private and public cloud resources do not change with time nor with the volume of demand;
- The other costs are either insignificant or do not depend on whether private or public cloud is used; and
- The demand for the resources must be served without a delay.

As opposite to [3], however, the data communication costs are not ignored in our model. As will be shown in the next section, their presence may have a significant effect on the overall costs and the optimal distribution between the private and the public cloud.

III. Estimating the Data Communication Costs in a Hybrid Cloud

In this section, the costs of open subsystems are estimated, and the effect of data communication on these costs is analyzed.

A. 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:

$$C = C_C + C_B, \tag{1}$$

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 private and public clouds:

$$\begin{aligned} C_C &= C_{CO} + C_{CC}, \\ C_B &= C_{BO} + C_{BC}, \end{aligned} \tag{2}$$

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.

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:

- $L_{BO} = k_O L_{CO}$ of data is transferred between the private cloud and the customers;
- $L_{BC} = k_C L_{CC}$ of data is transferred between the public cloud and the customers; and
- $L_{BOC} = \rho L_{BC} = \rho k_C L_{CC}$ of data is transferred between the organization and the public cloud.

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

$$L_{BO} + L_{BOC} = k_O L_{CO} + \rho k_C L_{CC} \tag{3}$$

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

$$L_{BC} + L_{BOC} = k_C L_{CC} + \rho k_C L_{CC}. \tag{4}$$

Now the costs can be rewritten as:

$$\begin{aligned} C &= p_{CO} L_{CO} + p_{CC} L_{CC} \\ &+ p_{BO}(k_O L_{CO} + \rho k_C L_{CC}) + p_{BC}(k_C L_{CC} + \rho k_C L_{CC}) \end{aligned} \tag{5}$$

Let us assume for simplicity that $k_O = k_C = k$, and that the unit prices of capacity in the private cloud is less or equally expensive than in the public clouds [13]. The higher unit price of a public cloud can be partly attributed the margins added by the cloud provider on top of its costs. Thus,

$$\begin{aligned} p_{CC} &= u p_{CO}, \\ p_{BC} &= u p_{BO}, \end{aligned} \tag{6}$$

where $u \geq 1$. Then, eq. (5) can be rewritten as

$$\begin{aligned} C &= p_{CO} L_{CO} + u p_{CO} L_{CC} + p_{BO}(k L_{CO} + \rho k L_{CC}) \\ &+ u p_{BO}(k L_{CC} + \rho k L_{CC}) \\ &= (p_{CO} + p_{BO} k) L_{CO} + \big(u p_{CO} + p_{BO} \rho k + u p_{BO} k(1+\rho)\big) L_{CC} \\ &= (p_{CO} + p_{BO} k) L_{CO} + \big[u p_{CO} + p_{BO} k\big(\rho + u(1+\rho)\big)\big] L_{CC} \\ &= (p_{CO} + p_{BO} k) L_{CO} + \big[u p_{CO} + p_{BO} k\big(u + \rho(1+u)\big)\big] L_{CC}. \end{aligned} \tag{7}$$

The unit prices p_{CO} and p_{BO}, as well as u can be seen as constants whose values are estimated by consulting public cloud providers' price lists (for public cloud) or by estimating the acquisition and operations costs over the depreciation period (for the private cloud).

The estimation of acquired capacity differs for the public and the private cloud. For the *private* cloud, the acquired capacity has a fixed cost whether or not it is used, and it represents the product of the maximum expected demand and the time. Let D denote the maximum demand for computing capacity, and let q denote the portion of that demand to be provided with the private cloud. Then, the acquired private cloud computing capacity is:

$$L_{CO} = qDT. \tag{8}$$

For the *public* cloud, on the other hand, the cost of acquired capacity is proportional to the amount of 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. Whereas the realistic demand curve may have multiple peaks, *for the purpose of the analysis it is rearranged, by sorting the data points in ascending order, to become 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:

$$d_C(t) = \begin{cases} 0, & \text{if } d(t) \leq qD, \\ d(t) - qD, & \text{otherwise.} \end{cases} \tag{9}$$

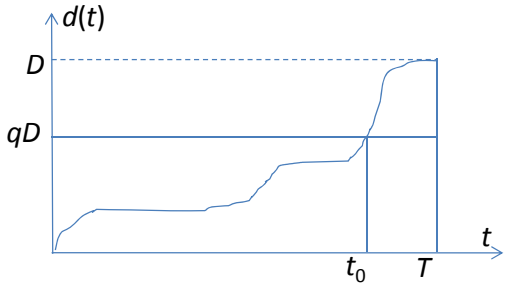

Figure 2. Demand curve rearranged to be monotonically non-decreasing

The acquired public cloud computing capacity can then be estimated as

$$L_{\mathrm{CC}} = \int_0^T d_{\mathrm{C}}(t)\,\mathrm{d}t = \int_{t_0}^T d_{\mathrm{C}}(t)\,\mathrm{d}t, \tag{10}$$

and the equation (7) can be rewritten as

$$C = (p_{\mathrm{CO}} + p_{\mathrm{BO}}k)\,qDT + \left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\big(u + \rho(1+u)\big)\right]\int_{t_0}^T d_{\mathrm{C}}(t)\,\mathrm{d}t. \tag{11}$$

B. Analyzing the Effect of Data Communication Costs

According to Equation (11) 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 communication occurs between the system and its customers, as reflected in the value of k; and on ii) how intensive interaction is needed between the private and public subsystems, as reflected in the value of ρ.

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:

$$\frac{\partial C}{\partial k} = p_{\mathrm{BO}}qDT + p_{\mathrm{BO}}\big(u + \rho(1+u)\big)\int_{t_0}^T d_{\mathrm{C}}(t)\,\mathrm{d}t > 0; \tag{12}$$

$$\frac{\partial C}{\partial \rho} = p_{\mathrm{BO}}k(1+u)\int_{t_0}^T d_{\mathrm{C}}(t)\,\mathrm{d}t > 0. \tag{13}$$

As could be seen, 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 ($t_0 < T$), the values of the partial derivatives in (12) and (13) are positive. Therefore, the costs increase as k and ρ values grow.

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 $u > 1$, then a hybrid cloud may have lower costs than the costs of purely private cloud or purely public cloud solution.*

Proof: Let us find the value of q minimizing the costs of the open subsystems. The partial derivative of C with respect to q is:

$$\frac{\partial C}{\partial q} = (p_{\mathrm{CO}} + p_{\mathrm{BO}}k)\,DT + \left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\big(u + \rho(1+u)\big)\right]\frac{\partial}{\partial q}L_{\mathrm{CC}}. \tag{14}$$

Let $t_{\mathrm{C}}(d)$ denote the inverse function of $d(t)$. Let us also define function $\tau_{\mathrm{C}}(d)$:

$$\tau_{\mathrm{C}}(d) = T - t_{\mathrm{C}}(d) \tag{15}$$

In fact, the value of $\tau_{\mathrm{C}}(d_0)$, where $d_0 = qD$, indicates the time of using the public cloud capacity given the value of q. Then, the acquired public cloud computing capacity L_{CC} can be evaluated by integrating over d:

$$L_{\mathrm{CC}} = \int_{d_0}^D \tau_{\mathrm{C}}(d)\,\mathrm{d}d. \tag{16}$$

Let $F(d)$ be an anti-derivative of $\tau_{\mathrm{C}}(d)$. Then,

$$L_{\mathrm{CC}} = \int_{d_0}^D \tau_{\mathrm{C}}(d)\,\mathrm{d}d = F(D) - F(d_0). \tag{17}$$

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

$$\frac{\partial}{\partial q}L_{\mathrm{CC}} = \frac{\partial}{\partial q}\Big(\int_{d_0}^D \tau_{\mathrm{C}}(d)\,\mathrm{d}d\Big) = \frac{\partial}{\partial q}F(D) - \frac{\partial}{\partial q}F(d_0) = -\frac{\partial}{\partial q}F(d_0) = -\frac{\partial F(d_0)}{\partial d}\frac{\partial d}{\partial q} = -\tau_{\mathrm{C}}(d_0)\,D. \tag{18}$$

Thus,

$$\frac{\partial C}{\partial q} = (p_{\mathrm{CO}} + p_{\mathrm{BO}}k)\,DT - \left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\big(u + \rho(1+u)\big)\right]\tau_{\mathrm{C}}(d_0)D. \tag{19}$$

The second derivative is:

$$\frac{\partial^2 C}{\partial q^2} = -\left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\big(u + \rho(1+u)\big)\right]D\,\frac{\partial}{\partial q}\tau_{\mathrm{C}}(d_0) = -\left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\big(u + \rho(1+u)\big)\right]D\,\frac{\partial \tau_{\mathrm{C}}(d_0)}{\partial d}\frac{\partial d}{\partial q} = -\left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\big(u + \rho(1+u)\big)\right]D^2\,\frac{\partial \tau_{\mathrm{C}}(d_0)}{\partial d}. \tag{20}$$

Recall that $t_{\mathrm{C}}(d)$ is inverse function of $d(t)$; furthermore, $d(t)$ is monotonically increasing. According to the inverse function theorem, for the domain where $d(t)$ is increasing, it holds that

$$\frac{\partial}{\partial d}t_{\mathrm{C}}(d) = \frac{1}{\frac{\partial}{\partial t}d(t)}. \tag{21}$$

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

$$\frac{\partial}{\partial d}\tau_{\mathrm{C}}(d_0) = \frac{\partial}{\partial d}\big(T - t_{\mathrm{C}}(d_0)\big) = -\frac{\partial}{\partial d}t_{\mathrm{C}}(d_0) < 0. \tag{22}$$

Thus, it follows that the second derivative is positive:

$$\frac{\partial^2 C}{\partial q^2} > 0. \tag{23}$$

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 $\frac{\partial C(q_{\min})}{\partial q}$ equals 0, then $q_{\min}$ minimizes C, i.e.

$$\frac{\partial C(q_{\min})}{\partial q} = (p_{\mathrm{CO}} + p_{\mathrm{BO}}k)\,DT - \left[up_{\mathrm{CO}} + p_{\mathrm{BO}}k\left(u + \rho(1+u)\right)\right]\tau_{\mathrm{C}}(d_0)D = 0. \quad (24)$$

Observing, that τ_{C} is also a function of q, we obtain:

$$\tau_{\mathrm{C}}(d_0; q_{\min}) = \frac{(p_{\mathrm{CO}}+p_{\mathrm{BO}}k)\,T}{up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)}. \quad (25)$$

By solving (25), the value of $q_{\min}$ can be found. If $u > 1$ and prices are positive, it follows that $0 < \frac{p_{\mathrm{CO}}+p_{\mathrm{BO}}k}{up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)} < 1$, and hence $0 < \tau_{\mathrm{C}}(d_0; q_{\min}) < T$. According to (22), $\tau_{\mathrm{C}}(q)$ is monotonically decreasing function in the domain $(0,1)$, and its values are within the region $(0,T)$. Therefore, there exists a value $q_{\min} \in (0,1)$ satisfying (25), i.e. a hybrid solution has lower costs than purely private cloud ($q = 1$) or purely public cloud ($q = 0$) solution, q.e.d.

Corollary. In the absence of data communication costs ($k = 0$), eq. (25) can be rewritten as:

$$\frac{\tau_{\mathrm{C}}(d_0; q_{\min})}{T} = \frac{p_{\mathrm{CO}}}{up_{\mathrm{CO}}} = \frac{1}{u}. \quad (26)$$

Thus, *the portion of the time when public cloud is used should be the inverse of the premium charged by the cloud software vendor*. This is in line with [3] 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 $1 - q_{\min}$ is the inverse of u. Indeed, for the uniformly distributed demand,

$$\tau_{\mathrm{C}}(d_0; q_{\min}) = T(1 - q_{\min}). \quad (27)$$

If $k = 0$, then the equation (25) simplifies to

$$T(1 - q_{\min}) = \frac{p_{\mathrm{CO}}T}{up_{\mathrm{CO}}}. \quad (28)$$

It follows that $1 - q_{\min} = {}^{1}/_{u}$, as in [3]. Note that according to this corollary, the regularity represented by (26) holds for the generic case of arbitrary monotonically increasing demand function, whereas only a special case of uniformly distributed demand was considered in [3].

Proposition 3: *The greater the data communication intensity of the service, as indicated by k and ρ, the more private cloud capacity should be acquired*.

Proof: Let $Q(\tau_{\mathrm{C}})$ be the inverse function of $\tau_{\mathrm{C}}(d_0; q)$, i.e.

$$q = Q(\tau_{\mathrm{C}}). \quad (29)$$

Recall that from (24), the value of q minimizing C can be found. By substituting (29) into (25) we can express the value of $q_{\min}$ as

$$q_{\min} = Q(\tau_{\mathrm{C}}) = Q\left(\frac{(p_{\mathrm{CO}}+p_{\mathrm{BO}}k)\,T}{up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)}\right). \quad (30)$$

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

$$\frac{\partial Q}{\partial k} = \frac{\partial Q}{\partial \tau_{\mathrm{C}}}\frac{\partial \tau_{\mathrm{C}}}{\partial k}. \quad (31)$$

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

$$\frac{\partial Q}{\partial \tau_{\mathrm{C}}} = \frac{1}{{\partial \tau_{\mathrm{C}}}/{\partial q}} = \frac{1}{\frac{\partial \tau_{\mathrm{C}}}{\partial d}\frac{\partial d}{\partial q}}. \quad (32)$$

Since $\frac{\partial \tau_{\mathrm{C}}}{\partial d} < 0$ (according to (22)) and since $\frac{\partial d}{\partial q} = D$, it follows that $\frac{\partial Q}{\partial \tau_{\mathrm{C}}} < 0$.

By taking partial derivatives from both sides of (25) we obtain:

$$\frac{\partial \tau_{\mathrm{C}}}{\partial k} = \frac{\partial}{\partial k}\left(\frac{(p_{\mathrm{CO}}+p_{\mathrm{BO}}k)\,T}{up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)}\right) = -\frac{p_{\mathrm{BO}}p_{\mathrm{CO}}T\rho(1+u)}{\left[up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)\right]^2} < 0. \quad (33)$$

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 (31) 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:

$$\frac{\partial Q}{\partial \rho} = \frac{\partial Q}{\partial \tau_{\mathrm{C}}}\frac{\partial \tau_{\mathrm{C}}}{\partial \rho}. \quad (34)$$

By taking partial derivatives from both sides of (25) we obtain:

$$\frac{\partial \tau_{\mathrm{C}}}{\partial \rho} = \frac{\partial}{\partial \rho}\left(\frac{(p_{\mathrm{CO}}+p_{\mathrm{BO}}k)\,T}{up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)}\right) = -\frac{(p_{\mathrm{CO}}+p_{\mathrm{BO}}k)\,T\,p_{\mathrm{BO}}k(1+u)}{\left[up_{\mathrm{CO}}+p_{\mathrm{BO}}k\left(u+\rho(1+u)\right)\right]^2} < 0. \quad (35)$$

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 (34) 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 portion of the capacity should be allocated to the private cloud, q.e.d.

IV. Numerical Experiments

In this section, some numerical examples are provided, wherein the effect of varying intensity of data communication in a hybrid cloud-based service is modeled. These examples are aimed at illustrating how the intensity of data communication affects the costs of open subsystems, and in particular how it affects the cost-optimal distribution of acquired capacity among the private and the public clouds.

An imaginary case of a hybrid cloud-based service is considered, 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 20 Amazon EC2 small instances [14], though this number may be changed without inflicting significant changes on the results of the experiments. A linear demand curve is assumed, i.e. the demand is uniformly distributed between zero and D as illustrated in Fig. 3. The use of linear demand curve, albeit unrealistic, enables easily finding the analytical solution to (25) and thereby helps in illustrating the proposed model; meanwhile, since the propositions in the previous section were shown to hold for an arbitrary monotonically non-decreasing demand distribution, the use of a more realistic demand distribution will not affect the results of the experiments.

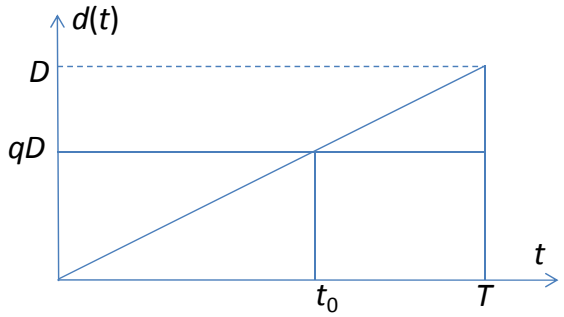

Figure 3. Uniformly distributed demand curve rearranged to be monotonically non-decreasing

For uniform demand, the cumulative acquired public cloud computing capacity is $L_{CC} = {}^{1}/_{2}\, DT(1-q)^2$. Therefore, eq. (11) can be rewritten as

$$C = (p_{CO} + p_{BO}k)\, qDT + \left[up_{CO} + p_{BO}k\left(u + \rho(1+u)\right)\right] \times {}^{1}/_{2}\, DT(1-q)^2. \quad (36)$$

The parameters in equation (36) are set the values as follows:

- One-month period is considered, i.e. $T = 24 \times 30 = 720$ (hours).
- The computing demand is assumed to be fully satisfied with 20 Amazon EC2 small instances, i.e. $D = 20$.
- 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.
- Price of public cloud computing capacity is estimated based on the price of standard small on-demand EC2 instance, located in EU, with Linux/UNIX): $p_{CC} = 0.095$ (USD/hour).
- Price of public cloud data transfer is estimated based on the "Data Transfer Out" pricing of EC2 for US & EU Regions, with total amount not exceeding 10TB/month: $p_{BC} = 0.15$ (USD/GB).
- The price of public cloud capacity is provisionally assumed twice more expensive than that of the private cloud [13], i.e. $u = 2$. Hence, $p_{CO} = 0.0475$ and $p_{BO} = 0.075$. Note that u can be changed without affecting the results, as long as $u > 1$.

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

In Fig. 4, the resulting costs of open subsystem are plotted as a function of the private cloud portion q. In the left part, the plots for different values of k are provided (the value of $\rho = 0.2$ is used). As could be seen, the costs grow as the value of k increases. 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 should be acquired. Furthermore, it can be shown that:

$$\text{as } k \to +\infty,\ q_{\min} \to 1 - \frac{1}{u+\rho(1+u)} = 0.6154. \quad (37)$$

The costs' dependency on the value of ρ depicted in the right part of the figure, exhibits a similar pattern. 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, the more the private cloud capacity should be acquired. It can be shown that:

$$\text{as } \rho \to +\infty,\ q_{\min} \to 1, \quad (38)$$

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 greater impact on the cost-optimal distribution of acquired capacity, than the communication between the open subsystems and the customers.

V. Summary and Concluding Remarks

Using a hybrid cloud, the organization's in-house computing capacity can be complemented with the computing capacity of a public cloud. In order to minimize the costs of such a hybrid cloud, a balance between the acquired private and public cloud capacity should be found: the higher price of the public cloud capacity should be compensated by the relatively short duration of the time, when the public cloud is utilized.

In this paper, the model for the hybrid cloud costs, encompassing the costs of computing and data communication, has been introduced. In the proposed model, the costs are modeled as a function of the portion of demand for computing capacity provided with the private cloud. Using the model, the

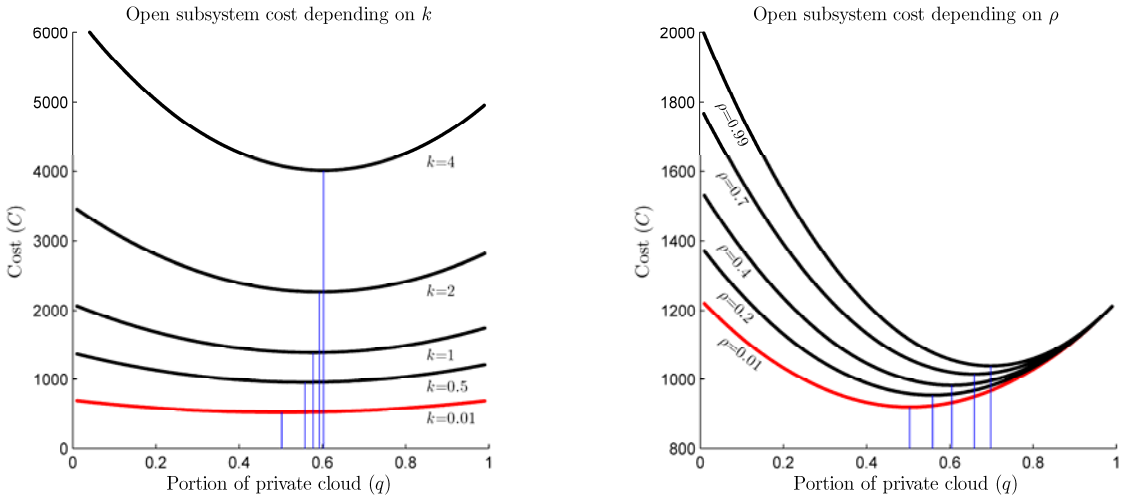

Figure 4. The cost of open subsystems for different values of k (left) and ρ (right), plotted as a function of the private cloud portion q. The vertical lines indicate the minimum costs for different values of k and ρ.

cost-optimal portion of private cloud computing capacity can be identified. Finding such optimal portion has been exemplified for the case of the demand uniformly distributed between zero and maximum levels.

By analyzing the model, it has been analytically shown that

i. The cost of open subsystems in the hybrid cloud increases with the intensity of data communication;
ii. A hybrid cloud may have lower costs than a purely private cloud or a purely public cloud solution has; and
iii. 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 should be acquired.
iv. Furthermore, it was also shown that, in the absence of data communication overheads, and given a 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.

Numerical experiments were used to illustrate the effect of data communication costs for the case, when the demand for computing capacity is distributed uniformly. As manifested in the experiments, the data transfer – either between the organization and the public cloud, or between the private/public cloud and the customers – increase the cost-optimal portion of computing capacity to be provided with the private cloud. Meanwhile, 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.

In summary, the introduced model contributes to the previous work in this domain by taking into account the data communication overheads when estimating the costs of a hybrid cloud, for a generic demand distribution function. In future work, this model can be extended by considering the price elasticity of computing and data transfer capacity. Other factors, such as trends in pricing and the net present value of investments, should be taken into account as well. Finally, in future work, the model should be complemented with the control cost incurred during the process of introducing the hybrid cloud into the organization.

ACKNOWLEDGMENT

The research reported in this paper was carried out in the frame 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

[1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing", Version 15, 10-7-09. National Institute of Standards and Technology, Information Technology Laboratory, available from http://www.csrc.nist.gov/groups/SNS/cloud-computing/, 2010 (last retrieved March 10, 2011).

[2] Youseff, L., Butrico, M., Da Silva, D.: Toward a Unified Ontology of Cloud Computing, Grid Computing Environments Workshop (GCE '08), pp. 1-10 (2008).

[3] J. Weinman, "Mathematical Proof of the Inevitability of Cloud Computing", Working paper, available from http://www.joeweinman.com/Resources/Joe_Weinman_Inevitability_Of_Cloud.pdf, January 8, 2011 (last retrieved on March 10, 2011).

[4] C. A. Lee, "A perspective on scientific cloud computing", Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC '10). ACM, New York, NY, USA, 2010, pp. 451-459.

[5] B. Pring, R.H. Brown, A. Frank, S. Hayward, and L. Leong, "Forecast: Sizing the Cloud; Understanding the Opportunities in Cloud Services". Gartner Dataquest, March 18, 2009.

[6] M. Risch and J. Altmann, "Cost Analysis of Current Grids and Its Implications for Future Grid Markets". In: Altmann J, Neumann D, Fahringer T (Eds.) Grid Economics and Business Models 5th International Workshop, GECON 2008, Las Palmas de Gran Canaria, 2008.

[7] J. Strebel and A. Stage, A, “An economic decision model for business software application deployment on hybrid Cloud environments”. In: Schumann, M.; Kolbe, L. M.; Breitner, M. H.; Frerichs, A. (eds.), Multikonferenz Wirtschaftsinformatik 2010. (Göttingen). Universitätsverlag Göttingen, 2010, pp. 195–206.

[8] M.E. Porter, “Competitive Strategy: Techniques for Analyzing Industries and Competitors”, Free Press, New York, 1980.

[9] J.B. Heide, “Plural Governance in Industrial Purchasing”. Journal of Marketing 67 (4), pp. 18-29, 2003.

[10] A. Parmigiani, “Why do firms both make and buy? An investigation of concurrent sourcing”. Strategic Management Journal 28 (3), pp. 285–311, 2007.

[11] N.P.Mols, “Economic explanations for concurrent sourcing”, Journal of Purchasing & Supply Management 16, pp. 61–69, 2010.

[12] P. Puranam, R. Gulati, and S. Bhattacharya, “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, September 2006 (last retrieved on March 10, 2011).

[13] A. Khajeh-Hosseini, D. Greenwood, J.W. Smith and I. Sommerville, “The Cloud Adoption Toolkit: Supporting Cloud Adoption Decisions in the Enterprise”. Software: Practice and Experience, 2011 (to appear).

[14] Amazon Elastic Compute Cloud (Amazon EC2), http://aws.amazon.com/ec2/.