Costs of Using Hybrid Cloud Infrastructure: Towards a General Framework

Oleksiy Mazhelis

University of Jyväskylä, Finland
oleksiy.mazhelis@jyu.fi

Abstract. Cloud computing infrastructure is a state-of-the-art computing as a utility paradigm, offering individuals and organizations instantly-available and scalable computing capacity. Organizations may deploy the cloud infrastructure in own data centers, as a private cloud, or use the public on-demand cloud infrastructure charged on a pay-per-use basis. The organizations may also adopt a hybrid solution, i.e. use public cloud capacity to complement the resources in the private cloud, e.g. during the periods of rapid growth in the demand. One of the important factors that affect the organizations' decisions to adopt a hybrid cloud is the total cost of acquiring and managing the infrastructure. In this paper, a general framework for cloud infrastructure cost assessment is introduced, wherein for several types of cloud infrastructure resources, the associated cost components and the factors determining these components are considered.

Key words: Cloud computing infrastructure, hybrid cloud, cost model.

1 Introduction

Cloud computing is a state-of-the-art computing as a utility paradigm, allowing the computing and storage capacity, as well as platforms and applications built on top of them to be provided to the customers on-demand in a scalable and efficient manner [5]. The lowest-level family of the cloud computing services is the so-called infrastructure as a service (IaaS) offering the customers the baseline computing, storage, and data communication capacities, while giving them the freedom to install and run on top of this infrastructure the applications of their choice.

Cloud infrastructure can be offered by a public cloud service provider (public cloud), and charged depending on the actual usage. Alternatively, the cloud infrastructure can be deployed as a so-called private cloud, i.e. within the organization's data center(s). Finally, the organization may combine both the in-house capacity of private cloud with the resources offered by the public cloud, to form a so-called hybrid cloud [5].

Cloud infrastructure services are adopted rapidly [6]; this rapid pace of adoption can be partly attributed to the cost savings for the customers promised by the cloud services [1]. On the other hand, available evidence suggests that the unit cost of public cloud capacity is higher than that in the private cloud [7, 3]. In other words, using public cloud resources only is likely to be more expensive in longer term than acquiring the needed resources up-front and managing them in-house.

A preprint version of the paper: Oleksiy Mazhelis, "Costs of Using Hybrid Cloud Infrastructure: Towards a General Framework", in M.A. Cusumano, B. Iyer, and N. Venkatraman (Eds.), Proc. of the 3rd International Conference on Software Business, ICSOB 2012, Springer LNBIP 114, pp. 261–266. The original publication is available at www.springerlink.com.

However, the use of expensive public cloud is economically justified when the need for computing capacity fluctuates. In this case, the use of the private cloud alone often results in over-provisioning, i.e. the infrastructure resources being underutilized most of the time. The hybrid cloud gives the opportunity to increase the utilization of the private cloud resources and hence minimize the overall infrastructure costs [8, 4].

The cost of the cloud infrastructure was studied in several works, where a number of factors determining whether the hybrid solution brings cost savings were identified:

- The degree of demand fluctuation [8, 4] and demand growth predictability [9];
- The pricing models applied to the private and public resources [8, 4];
- The communications overheads and the effect of volume discounts in the above pricing [4], as well as the expected trends in pricing [7, 3];
- The net present value of money related to the expected lifetime of solution [7, 3];
- The start-up costs and/or the costs of transforming the current solution towards enabling a hybrid solution [3].

As could be seen, the state-of-the-art research on the hybrid cloud costs is rather fragmented, with different research efforts dealing with individual aspects of the issue. In this paper, we aim at elaborating a generic costs framework for hybrid cloud infrastructure, where these factors are integrated. In the framework, several cost components are considered, including the cost of computing, data communications, and data storage. These cost components are categorized according to their likely effect on the overall costs, which in turn depends on whether the cost component is affected by the interaction between the private and the public clouds.

The remainder of the paper is organized as follows. In section 2, the cost components and the factors affecting them are considered. The identified cost components are classified in section 3 into three categories. Finally, section 4 concludes the paper and outlines the directions for further work.

2 Hybrid cloud infrastructure cost

In this section, we decompose the costs of hybrid cloud infrastructure into cost components based on the resource whose usage incurs these costs, and consider different factors that may have an effect on the identified cost components.

An organization can allocate the workload to the private and public portions of a hybrid cloud in two ways. In case the organization deals with the workload of heterogeneous nature, the allocation of workload to the infrastructure can be based on the workload type: for instance, the tasks involving sensitive data or having carrier-grade performance requirements may be assigned to the private infrastructure, while less critical tasks tolerating occasional delays may be allocated to the public infrastructure. Alternatively, in case the workload is homogeneous, the organization may decide on where to allocate the workload on the fly, depending on the current load - e.g. by offloading the peak load exceeding the private capacity to the public cloud, and using the private infrastructure otherwise. For simplicity, in this paper we will assume the presence of homogeneous workload and hence the second type of allocation.

The cost of the hybrid cloud infrastructure can be decomposed into the cost components according to the capacity provided by the cloud and charged for. Taking the Amazon Elastic Compute Cloud (EC2, `http://aws.amazon.com/ec2/`) as an example of the cloud infrastructure offering, the cost components include the costs of computing, persistent storage, data communications, load balancing, and monitoring. The cost incurred by each component is determined by the usage patterns, the charging scheme applied, and other factors considered below.

Dependency on the demand/usage. The cost of an infrastructure capacity can be estimated as a product of i) the volume of the capacity charged for and ii) the unit cost of that capacity (taking into account the time of expected usage as well as possible reservation charges and volume discounts).

The volume of the capacity charged for depends on whether the capacity belongs to the private or to the public portions of the cloud. In the private cloud, the capacity is usually acquired up-front, configured and integrated, and then operated and maintained throughout its lifecycle. Its cost is proportional to the amount of resources acquired, and hence is proportional to the maximum demand it serves. The private cloud costs are rather independent of the actual degree of utilization (in fact, power consumption is affected by the computing load, but the effect is not dramatic [2]).

The capacity charged in the public cloud depends on how much the capacity is used. Its consumption can be measured on hourly or monthly basis, based on counting the number of virtual instances (e.g. computing capacity), on counting the volume of data (data communications and storage), or on whether the capacity is used or not during the period of interest (e.g. load balancing). As could be seen, the cost of public cloud infrastructure depends on the capacity usage patterns. The computing and data communication capacity are taken into use upon need and released as soon as the computing task or data communication is completed; the capacity usage is therefore likely to follow the peaks and drops in demand. On the other hand, the persistent storage capacity, once used, is likely to remain used for a long period of time; hence, the capacity usage is likely to exhibit eventual growth rather than fluctuations.

Time dimension. The continous growth of the computational power of hardware is likely to result in eventual decline of the prices for infrastructural resources. Indeed, Amazon cut the prices of on-demand instances by 15% in November 2009[1] and the data transfer prices by 20% in July 2011[2]. Therefore, some researchers assume the prices of public cloud infrastructure resources to decline by 15% on a bi-annual basis [3].

Similarly, the equipment procured for the private cloud is subject to price reductions, due to accelerating price-performance ratios in computing power (expressed by Moore's law), bandwidth availability (expressed by Gilder's law), and storage capacity (the GB-per-dollar ratio has been doubling every 14 months[3]).

[1] http://aws.amazon.com/about-aws/whats-new/2009/10/27/announcing-lower-amazon-ec2-instance-pricing/

[2] http://aws.amazon.com/about-aws/whats-new/2011/06/30/aws-announces-new-data-transfer-pricing/

[3] http://www.mkomo.com/cost-per-gigabyte

Besides the pricing trends, the time value of money should be taken into account when assessing the costs of both the private and the public cloud infrastructure. In particular, the net present value (NPV) of cash flow over time can be estimated [7]. Generally, the NPV analysis favors the public cloud option, due to the decaying factor applied to the future expenses as compared with the up-front costs.

3 Types of cost components

The contribution of the cost components above to the overall costs and their effect on the cost-efficient division between the private and the public cloud infrastructure depends on whether a cost is incurred due to the interaction between the private and public portions of the private cloud. Therefore, in this section, the costs components are divided into the categories of: i) the constant costs due to the adoption of a hybrid infrastructure, ii) the costs depending on the usage of the private or public portions of the hybrid infrastructure, and iii) the costs depending on the interaction between the private and public portions of the cloud, which are considered separately below.

Constant costs due to the adoption of a hybrid infrastructure. Belonging to this category are the invariable costs incurred due to the adoption of cloud, whose value is rather independent of the intensity of use - and hence independent of the specific division between the private and the public clouds. These costs can be exemplified with the costs of Amazon EC2 elastic load balancing or detailed monitoring service, which are charged independently on how intensely the infrastructure is used. These cost components depend partly on the maximum expected demand and the specifics of the service: for instance, given the maximum demand below a certain limit and best effort service quality guarantees, the elastic load balancing may be unnecessary.

Being independent on the usage, these costs do not affect the optimal division between the public and private cloud. Still, the presence of these costs may rise the cost of the hybrid cloud infrastructure, thereby potentially making the private or public cloud the cost-efficient solution. Thus, taking these costs into account is important.

Costs incurred due to using private or public cloud infrastructure. The costs of the components in this category, exemplified with the cost of computing capacity, vary depending on how intensively they are used. More specifically, with respect to the private portion of the infrastructure, these costs depend on the amount of capacity acquired up-front, and therefore depend on the maximum demand that the private infrastructure is expected to serve. With respect to the public portion of the infrastructure, the costs depend on the amount of capacity that has been consumed over the charging period.

These costs are affected by the split of the load between the private and the public portions of the infrastructure: the smaller the threshold demand served with the private infrastructure, the less the amount of equipment to acquire and operate in-house, the greater the portion of public infrastructure that needs to be provisioned on demand. The cost-efficient division between the private and the public portions of the infrastructure is achieved when the time of using the public infrastructure is inversely proportional to the premium charged by the public infrastructure service provider. It has two implications:

i) the greater the premium, the greater the portion of private cloud infrastructure, and ii) the greater the fluctuation of the demand, the greater the cost benefit of the hybrid cloud as compared with the fully private cloud infrastructure [8, 4].

It should be noted that, whereas the computing resources are released as soon as computing task is completed, the demand for persistent storage capacity often accumulates over time. In this scenario, the storage consumption can be approximated as a monotonically increasing function of time (compare with the fluctuating demand for the computing capacity). Thus, assuming that the overall volume of storage can be predicted correctly, the cost of storage depends on the cumulative storage capacity consumed - i.e. effectively on the actual usage. It can be shown that in this case, assuming that the private and the public cloud storage capacities are acquired and charged with the same interval (e.g. monthly) and have the same unit prices, the total storage cost in the hybrid cloud stays constant independently on how storage is distributed between the private and the public infrastructure. As a result, the cost-efficiency of private vs. public storage in this case is determined by the pricing of the private and public infrastructure (including volume discounts), and the intervals between storage acquisitions.

Costs depending on the interaction between the private and public cloud. These costs depend not only on the intensity of using the private and the public infrastructure, but also on the intensity of interaction between the two. Such interaction affects, e.g., the costs of data communications and the costs of persistent data storage.

Consider first the data communication costs. Let us assume that the intensity of the interaction between the private and the public portions of the hybrid infrastructure is reflected in the volume of data transfered between them. It was found that, the greater the intensity of interaction, the greater the private portion of the hybrid infrastructure that should be acquired in order to minimize the overall costs [4].

The use of persistent storage may as well incur an interaction between the private and the public clouds. Two illustrative scenarios can be envisioned:

– *No interaction.* The service-related data is persistently stored by the private and the public clouds independently, with no replication or synchronization between the two. In this scenario, the storage costs will be incurred by the private and the public clouds independently, and will therefore belong to the second cost category.
– *Intense interaction.* The data is stored in the private cloud, and a full replica is stored also in a remote public cloud, to mitigate the risk of losing the data if stored in a single physical location. In this scenario, the interaction is rather intense, and hence the storage costs belong to the third category.

4 Conclusions

In the previous sections, the generic framework for the costs of using a hybrid cloud infrastructure has been introduced. The framework accounts for various cost components, including the costs of computing capacity, persistent storage, data communications, load balancing, and monitoring. A number of factors affecting these cost components have been identified including demand fluctuation and dynamics, the unit costs and demand elasticity of the unit prices, as well as the evolution of the above factors in time.

Based on the possible effects of these factors, the cost components are classified as the constant costs, the costs depending on the usage of private or public clouds, or the costs depending on the interaction between the private and public portions of the cloud.

Among the factors affecting the costs, only the workload division between the private and the public portions of the cloud infrastructure is controlled by the organization using the cloud services, whereas the others are external variables mainly determined by the end-customers (shaping the demand) or by the cloud service providers. Therefore, the organizations using cloud infrastructure services can seek a cost-efficient solution by adjusting the workload division between the private and the public clouds.

Due to the effect of fluctuating demand, the costs of computing capacity and data communications are often minimized by using the hybrid cloud. Meanwhile, as the demand for persistent storage usually accumulates over time, the storage costs are at their minimum in case the private cloud only is used. Furthermore, since the overall storage cost is relatively independent on how the storage is distributed between the private and the public clouds, as compared with the costs of computing capacity and data communications costs, the latter largely determine the shape of the total cost function. As a result, the minimum of total cost is usually achieved by using a hybrid solution.

In future work, the proposed framework shall be extended by taking into account the growth of customer demand, caused either by the growth of customer base or by intensified usage of individual customers. The framework shall be also validated by applying it to analyzing the costs of hybrid cloud infrastructure in real-world scenarios.

References

1. CIO MAGAZINE. Cloud computing survey. Tech. rep., CIO magazine, Available from http://www.cio.com/documents/whitepapers/CIOCloudComputingSurveyJune2009V3.pdf, June 2009.
2. GREENBERG, A., HAMILTON, J., MALTZ, D. A., AND PATEL, P. The cost of a cloud: research problems in data center networks. *SIGCOMM Comput. Commun. Rev. 39* (December 2008), 68–73.
3. KHAJEH-HOSSEINI, A., GREENWOOD, D., SMITH, J. W., AND SOMMERVILLE, I. The cloud adoption toolkit: supporting cloud adoption decisions in the enterprise. *Software: Practice and Experience 42* (2011), 447–465.
4. MAZHELIS, O., AND TYRVINEN, P. Economic aspects of hybrid cloud infrastructure: User organization perspective. *Information Systems Frontiers* (2011), 1–25.
5. MELL, P., AND GRANCE, T. 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/, 2010.
6. PRING, B., BROWN, R., FRANK, A., HAYWARD, S., AND LEONG, L. Forecast: Sizing the cloud; understanding the opportunities in cloud services. Gartner dataquest, March 18 2009.
7. WALKER, E. The real cost of a cpu hour. *Computer 42* (2009), 35–41.
8. WEINMAN, J. 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 October 28, 2011), January 8 2011.
9. WEINMAN, J. Time is money: The value of "on-demand". Working paper, available from http://www.joeweinman.com/Resources/Joe_Weinman_Time_Is_Money.pdf (last retrieved on October 28, 2011), January 7 2011.