A Cost-Effective Methodology Applied to Videoconference Services over Hybrid Clouds

Javier Cerviño · Pedro Rodríguez · Irena Trajkovska · Fernando Escribano · Joaquín Salvachúa

Abstract This paper tackles the optimization of applications in multi-provider hybrid cloud scenarios from an economic point of view. In these scenarios the great majority of solutions offer the automatic allocation of resources on different cloud providers based on their current prices. However our approach is intended to introduce a novel solution by making maximum use of divide and rule. This paper describes a methodology to create cost aware cloud applications that can be broken down into the three most important components in cloud infrastructures: computation, network and storage. A real videoconference system has been modified in order to evaluate this idea with both theoretical and empirical experiments. This system has become a widely used tool in several national and European projects for e-learning and collaboration purposes.

Keywords videoconference · cloud computing · hybrid clouds · cost-effectiveness analysis

1 Introduction

“Divide et impera”. Divide and rule. This maxim was first used by the Roman general Julius Caesar and the French emperor Napoleon. The first large-scale application of this rule was in 168 BC, when the Romans divided Macedonia into four independent republics to govern each of them easily. Recently, in computer science it is mostly known as the Divide and Conquer algorithm. This works by recursively separating a computer problem into many sub-problems, until these become affordable enough to be solved directly. In this paper we have based on Cloud technologies to introduce a new perspective to this long-lived proverb that is mainly related to hybrid clouds and cost-effective strategies.

Over time hybrid cloud has proven to be a valid solution for the business sector as many companies that were initially avoiding public cloud solutions, later showed higher confidence in the hybrid model as a feasible use case solution of the cloud computing model itself. Cloud users and the companies in particular, accepted leaving private IT assets inside and migrated less sensitive data and operations to the public cloud.

Overcoming the heterogeneity of IaaS billing policies and working out the best combination of public-private and cross-cloud infrastructures is a tempting challenge. In our research, we start from here in order to give another reason to deploy services in a hybrid cloud: to efficiently enhance the use of resources. There are many systems that can benefit from deploying on hybrid clouds instead of only using a single cloud. We provide guidelines for developers to design their applications and services according to their requirements.

We want to validate this concept by means of a system that offers videoconference to users on both private and public clouds. We have designed, developed and tested a new architecture for a session-based videoconference system where several users can join and control (schedule, delete and modify videoconference) sessions through a web application. The system focuses on optimal use of the available resources. To that extent, we studied the existing hybrid infrastructures that were used for purposes other than videoconference and we based our work on similar videoconference systems on various Cloud infrastructures.

This paper is organized as follows: in Section 2 we talk about the state of the art in the domain of hybrid cloud infrastructures. Section 3 introduces the main motivation of this research, explaining how hybrid architectures can show better performance in some scenarios. Section 4 goes through the validation of a videoconference system in terms of cost and resource use. Section 5 presents a real case scenario as a practical validation of the hybrid system and cost formulae, numbering the projects and researches using the Conference Manager. Finally in Section 6, we conclude our work by encouraging other researches to try our outcomes for further research.

2 State of the art

In this section we are going to present the literature review divided into two parts: first one presents existing videoconferencing systems and in second one we revise current methodologies for deployment in cloud and cross cloud techniques and architectures.

2.1 Videoconference systems

There are systems in the domain of videoconference that allow users to schedule web videoconference sessions or participate through their web browsers. These are for instance, FlashMeeting, Adobe Connect, WebEx, GoToMeeting, Skype, etc. Table I in [2] classifies the characteristic features offered by each of these systems.

Yet another example [12], presents a prototype for a conferencing system in the cloud based on the SOA approach. Although this framework has very similar features as our system, such as a Conference Management component, it still lacks of a complete implementation and verification. Moreover initially the system is not thought to operate over heterogeneous or hybrid clouds.

2.2 Methodologies for deployment in the cloud

With the objective of quick responsiveness to the business challenges, hybrid solutions in the cloud have shown to be an inevitable approach especially common among the industry-specific applications [6, 7]. To avoid risky undertakings migrating entire systems on the cloud, some companies commit themselves to the hybrid approach, which, as the literature states, has remarkably increased profitability.

We located similar research in the area of cost-optimization techniques for hybrid clouds. For example in [5] they focus on scheduling deadline constrained workloads with a minimum execution cost on a cloud by following the application's QoS requirements such as CPU and network. Song et al. [13] present an approach following similar idea as ours for optimized cloud provider selection, but they do not introduce a division of bandwidth, CPU and storage sensitive components. Other research like [11] present an optimization approach for profit maximization on cloud. This method is aimed at applications running on one cloud and the cost optimization is conditioned by QoS and SLAs. In our technique on the other hand, we are guided by price constraints across multiple clouds. And in [10] authors show that cloud price and server bandwidth play the most important roles in saving cost, i.e. such hybrid model can save up to 30% bandwidth expense compared with the Clients/Server mode. Furthermore, Li et al. [9] propose a measurement tool for comparing four major Cloud providers in order to select the best-performing provider for a given application of a Cloud customer.

While these methods try to define only one best matching cloud, we intend to establish a general methodology to enable cost planning for system deployment on multiple provider hybrid clouds.

3 Motivation and context

A videoconference system that allows a great number of users per conference, multiple simultaneous conferences, different client software (requiring transcoding of audio and video flows) and provides an automatic recording service, as the one we have built requires a lot of computing resources. Typical videoconferencing scenario (like the one explained in [3]) includes several videoconference clients. Some are connected through a Multipoint Control Unit (MCU) and others participate via Flash or SIP. In both cases transcoding the data flows is necessary. The scenario also includes a Real Time Messaging Protocol (RTMP) server for the flash clients and a SIP server for the SIP clients.

In order to allow a cost-effective scaling of our videoconference system, the use of cloud computing resources appears as a natural approach, since they provide an illusion of infinite computing resources available on demand and the ability to pay for use of computing resources on a short-term basis as needed [1].

However the use of cloud computing resources from a single provider comes with several disadvantages as shown in [1, 8]. Critical problems that can benefit from hybrid cloud architectures are: Geographical location and legal issues, cost and lock-in, service availability, wasting of existing resources in private clouds and security.

In light of the problems listed above, the use of resources from different providers as well as private resources can help us to provide a service with better performance, lower cost and to avoid or at least mitigate most of the problems of cloud computing. This will applied to our videoconference service in Section 4 of this paper.

To be able to effectively exploit the hybrid clouds two things are required. First we need to make use of a virtual infrastructure manager [14] to provide a uniform and homogeneous view of virtualized resources, regardless of the underlying virtualization platform. Second, we need to split our service into three parts:

- *CPU intensive modules*. Parts of the application that consume most of the CPU cycles needed to provide a service. In our case we have identified the transcoding and recording modules of our videoconference system as the CPU intensive modules.
- *Bandwidth intensive modules*. These are modules that consume most of the bandwidth. In our videoconference system, the MCUs and RTMP servers are bandwidth intensive components.
- *Storage intensive modules*. Disk servers and databases fall into this category. In our case the recorded conferences are stored in a NFS server.

This division gives us the opportunity to place the modules that need a specific kind of resource where it better serves our needs and objectives. We have named this partition *Cloud computing Resource Oriented Partition* or CROP.

4 Validation of hybrid cloud for a videoconference system

This section introduces a general methodology to calculate traditional Cloud-node costs, based on the principles explained in Section 3. An example of a videoconference system (Section 4.2) in order to put into practice the previous methodology (Section 4.3) is presented afterwards.

4.1 General methodology

Figure 1 represents the typical costs of a node hosted in a particular Cloud infrastructure: computation time, traffic data and storage cost.

The next cost-calculating formulae are completely based on the way Amazon is charging its clients. Although almost all of its competitors (like RackSpace, GoGrid, Azureus, ...) offer their own payment methods, they basically follow Amazon's model. Others for instance offer AWS-compatible cost calculators.

We will start explaining the cost formulae by denoting X_i as the representation of a Cloud provider. Let's assume we want to contract the services of a provider, which is offering low-prices in the use of CPU (named X_{cpu}), other provider that offers better prices in bandwidth consumption (named X_{bw}) and a third provider, which has special deals in storage services (named X_{stor}).

Then, given the architecture explained before and basing on the Cloud Price Calculator we mentioned in Section 1 we can work out the cost of this architecture in a hybrid cloud environment.

$$C_{\text{total}} = \sum_{i \in \{bw, cpu, stor\}} C_{cpu}(X_i) + C_{bw}(X_i) + C_{\text{stor}}(X_i) \tag{1}$$

This formula shows the sum of the three aforementioned services: C_{cpu},C_{bw} and C_{stor} for each Cloud provider (X_i). Each of these components is further detailed as follows:

$$C_{cpu}(X_i) = \sum_j C_{cpu}(j, X_i)\, t \tag{2}$$

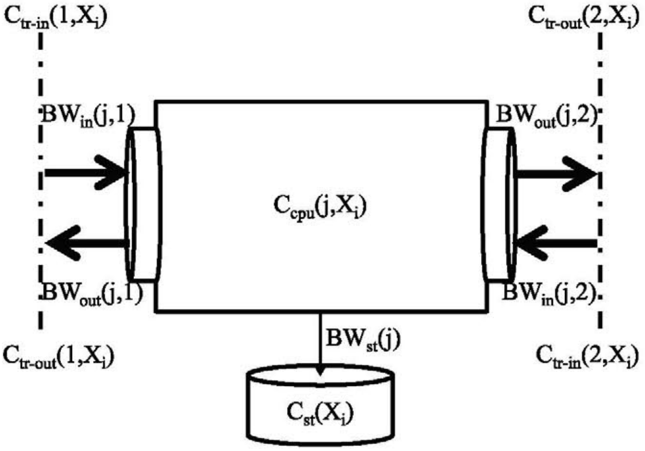

$C_{tr\text{-}in}(1,X_i)$
$C_{tr\text{-}out}(2,X_i)$
$BW_{in}(j,1)$
$BW_{out}(j,2)$
$C_{cpu}(j,X_i)$
$BW_{out}(j,1)$
$BW_{in}(j,2)$
$C_{tr\text{-}out}(1,X_i)$
$BW_{st}(j)$
$C_{tr\text{-}in}(2,X_i)$
$C_{st}(X_i)$

Fig. 1 Typical Cloud node costs

This formula along with the following ones complies with Fig. 1, in which we can see all the components that are part of each node cost. Formula 2 gives us the total computation cost for each cloud. We have to sum up the computation costs of each j node that is running on this X_i provider and the computation capacity it uses. For example, in the case of a Web server we will need a medium level CPU while in the case of a video transcoder we will need more computation capacity. Given that there are some providers that charge on memory capacity for each virtual machine, the reader could include this cost as part of the $C_{cpu}(j, X_i)$ value. Finally, t is the number of hours that are going to be considered.

$$C_{bw}(X_i) = 3600\,t \sum_{j} \sum_{k} C_{\text{int}}(j, k, X_i) \tag{3}$$

Formula 3 is the result of adding together all costs generated from different traffic sources in each j node of X_i Cloud provider. In this case we assume the constant 3600 because traffic is measured in bytes per second, and here the costs are per hour.

In this case we have simplified the figure by representing only two network interfaces, one on the left-hand side (numbered 1) and other on the right-hand side (numbered 2). However the formula takes into account the total number of interfaces that are attached to the node, and each of them are denoted by k. The resulting traffic cost is the sum of all cost interfaces. We want to clarify that here we are referring to the node interfaces and not to the Cloud interfaces, so the reader could consider traffic that is going to be sent out of the Cloud datacenter, and traffic that is going to be sent to other machines in the same datacenter. Both communications could have different costs because the former is considered as external transfer of the Cloud network, and the latter is part of the internal network transfers.

$$\begin{aligned} C_{\text{int}}(j, k, X_i) = {} & C_{\text{tr-in}}(k, X_i)\; BW_{\text{in}}(j, k) \\ & + C_{\text{tr-out}}(k, X_i)\; BW_{\text{out}}(j, k) \end{aligned} \tag{4}$$

In formula 4 each interface has incoming traffic (BW_{in}) and outgoing traffic (BW_{out}) with their corresponding cloud costs: $C_{\text{tr-in}}$ for incoming traffic and $C_{\text{tr-out}}$ for outgoing traffic. Both traffic components need to be measured in bytes per second. The sum is the cost of each interface.

$$C_{\text{stor}}(X_i) = \sum_{j} 3600\,t\, C_{st}(X_i)\; BW_{st}(j) \tag{5}$$

Finally, formula 5 is related to the storage costs of each j node of X_i Cloud provider. Here we have defined the storage as a constant data flow saved on virtual disks. Next, we have introduced BW_{st}, as the rate (per second) at which we store data on the disk.

4.2 Simple videoconference system

In [3] we decided to compare two topologies: a system in which all resources were allocated in the same public cloud, and another system in which this allocation was made by using a public Cloud and our own datacenter. However this study does not exactly coincide with the current idea of cost calculation because this service is supposed to take advantage of using several public clouds where we have to pay for almost everything. For this reason in the current work the problem will be tackled through a different approach.

In order to better validate our methodology we have designed a traditional videoconference system focused on offering the service to multiple participants who want to join in a meeting session. This system is a simplified service of the one explained in [3]. From now onwards, we will consider the total storage cost is zero because we use our own datacenter to store the recorded videos.

Here we have only taken into account three essential components that present features such as video and audio communication among users, real-time streaming and recording of the session. First component: the *Web Flow Server*, is responsible for forwarding all the traffic between the transcoder and each of the users. Second component: the *Transcoder*, composes a video of the entire session. This video usually consists of 1–5 videos (each of them from different users) and is coded in H.264 format. Note that the transcoder generates a video with a resolution of 1024 × 768 pixels. Regarding the audio streams, the transcoder is responsible for joining all streams into one to be used for recording by the recorder component. Last component: the *Recorder*, stores the video and audio streams. The configuration of this system stores video files in MP4 format, with the video and audio generated in the transcoder.

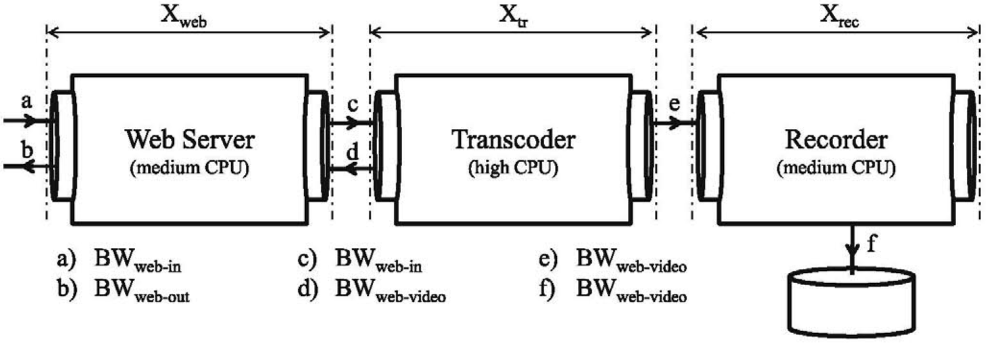
In Fig. 2 we can see the architecture with the three elements interconnected between themselves, sending session media streams.

$$BW_{\text{web-in}} = BW_{\text{web-user}} \min\{N_{\text{web-users}}; N_{\max}\} \tag{6}$$

$$BW_{\text{web-out}} = BW_{\text{web-video}}\; N_{\text{web-users}} \tag{7}$$

Formula 6 calculates the total incoming bandwidth consumption in the external system interface by multiplying the video and audio stream bandwidth from the user ($BW_{\text{web-user}}$) by the number of users that appear

Fig. 2 Cloud architecture



in the generated video. This number is considered to have a limit of $N_{\max}$ users, so we have to take the minimum value between this limit and the number of connected web users ($N_{\text{web}-\text{users}}$). Formula 7 refers to the outgoing bandwidth, that is the amount of bandwidth consumed by the video and audio streams generated in the transcoder. These streams are sent to each web user ($BW_{\text{web}-\text{video}}$).

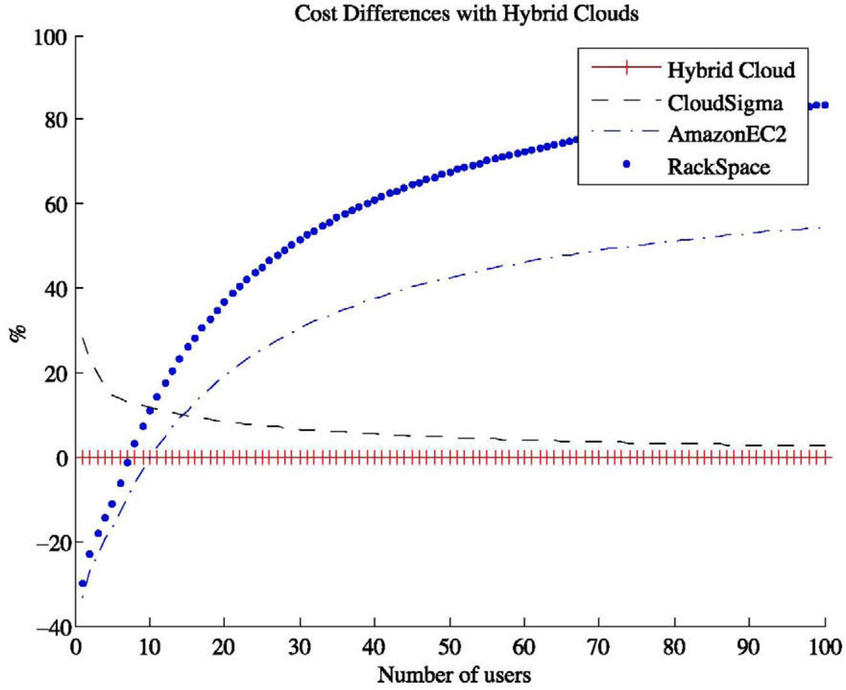
4.3 Cost analysis

In our research we have decided to analyze four scenarios with real Cloud providers: Amazon AWS, CloudSigma and Rackspace. Given their published prices we can consider them as low-priced Cloud providers for CPU, bandwidth and storage, respectively. In our first scenario (named hybrid scenario) the Web server is hosted in CloudSigma data centers, while the Transcoder is instantiated in Amazon AWS Cloud and the Recorder is in Rackspace public Cloud. The different Clouds were connected through their public interfaces, using public IP address in all components. In the rest of the scenarios all the components are in only one of these Cloud providers. For the results we have taken into account that in single Cloud scenarios the traffic must be considered internal, which in general presents lower prices.

For calculation we can replace the values of bandwidth streams (BW_{in}, BW_{out}, etc.), Fig. 1, with the corresponding ones in scenarios represented on Fig. 2.

Figure 3 depicts the cost comparison for each scenario. We have considered the difference in cost between each scenario and the hybrid scenario when we increase the number of videoconference participants. Therefore, each curve above 0 means that the referenced scenario is more expensive than the hybrid one.

Fig. 3 Cost differences between Cloud architectures



In our example we can see that for more than 10 people we need to use this hybrid architecture in order to obtain a cost-effective service. On the other hand, for less people we can find other traditional Cloud solutions. Another interesting situation occurs when the number of users connected to the videoconference increases to above 60 people. In this situation we can see that the cost of the CloudSigma scenario approaches the hybrid scenario. This happens because the bandwidth consumption turns into the most expensive factor in the formula, and in this case both the hybrid and the CloudSigma scenarios have the same bandwidth values.

5 Results validation and performance evaluation

In this section we present the test scenarios we have used in order to validate the formulae established in the previous section and show the outcomes of the validation. Afterwards we number the projects in which our system was used and finally we trace the way for future research directions that can be motivated from our work.

5.1 Test scenarios

We have established five test scenarios based on the architecture explained in [3], all of them including six participants in a videoconference sessions. This architecture is based on Isabel, a videoconference tool which supports sessions with multiple participants, and on Conference Manager, that schedules Isabel videoconference sessions and provides additional services, such as video recording. Five of these tests are connected through the Isabel application and one via web client portal. The scenarios are differentiated by the videoconference mode. The tests were set up to one hour in duration. We hosted a Transcoder and an Isabel Flow Server in two medium Amazon EC2 instances, while a Web Flow Server was hosted in our private datacenter.

In the first scenario we set-up an Isabel *chat mode*, meaning various videos are displayed simultaneously on a single screen. In the second and third scenarios all the clients were set up in *VNC mode* for displaying a video and some slides, respectively, using the VNC technology. Scenarios 4 and 5 replaced the *VNC mode* with a *VGA mode* to show video and slide presentations similarly as the scenarios 2 and 3. In this mode there is a video display shared among all the users and they can view video, applications, etc.

The video displayed is obtained frame by frame from a screen, encoded and sent. The use of VGA video has significantly decreased the cost for the outgoing traffic as compared to VNC mode with video. The key is the fixed transmission rate of the VGA video that does not necessarily use the entire bandwidth for transmission.

The results from Fig. 4 indicate an appealing cost-performance trade-off for the hybrid videoconference system. We can see that even the use of VNC in the conference session has not influenced much on increasing the total cost of the incoming and outgoing traffic, which remained within reasonable limits. This

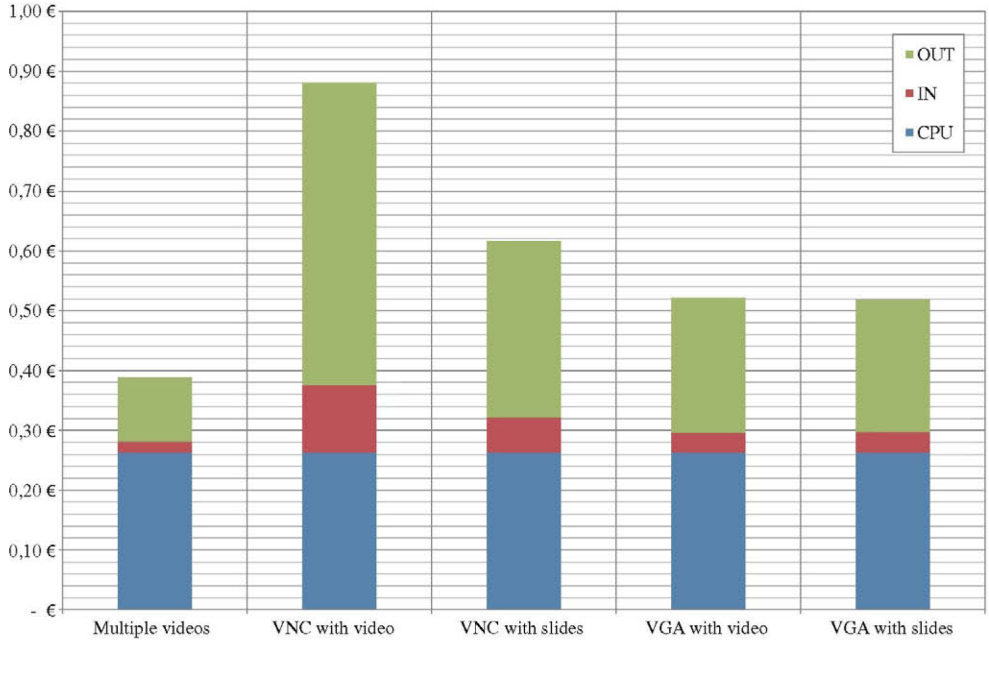


Fig. 4 Cost of an Isabel Hybrid Session

has confirmed our expectations that a hybrid model could be a good solution for hosting videoconference systems that require an optimized cost policy and good quality of service.

5.2 Project resource usage

The Conference Manager has been successfully integrated into several projects allowing us to obtain further proof that the solution is valid. The most important of these projects are Global Project, in which it was used in meetings of TERENA, EGEE, W3C among others; GATE for the organization of classes (from five to ten two-hour classes are scheduled every week); and CyberAula 2.0 to record, store and stream courses from different Universities.

At the time of writing this paper, the system had been used to organize 592 sessions with 941 videos stored.

5.3 Research using conference manager

As we mentioned in Section 3, using hybrid clouds can be useful because of the geographical diversity of the offer. In our experience, intercontinental videoconferencing has proven to be a challenge, especially when relying on TCP as the transport protocol as it is the case of our web clients. As seen in [4], the network infrastructure used by typical commercial clouds usually performs really well, even between different continents giving users a good service in terms of packet loss, delay and throughput among the nodes within the cloud. In a hybrid case as the one presented in this paper, we can take into account the variety of locations provided by the different providers and the geographical location of the service's users and choose the provider that suits us best.

6 Concluding remarks and open challenges

In this paper, we have presented a cost-effective methodology for developing and deploying applications over multi-provider hybrid cloud. The core idea of this methodology is to divide the application into three parts: CPU, bandwidth and storage intensive modules. Whenever this is possible, this methodology aims to optimize costs by deploying each of these modules in the most suitable cloud provider. We have validated this methodology in our videoconference system. Firstly, we introduce guidelines to calculate traditional Cloud node costs and apply it to the videoconference scenario concluding that the proposed deployment strategy does reduce costs on paper. To confirm this theoretical validation, we have successfully tested the methodology in a real videoconference environment.

We would encourage those developers who are implementing both services or cloud middlewares to take our work into account, because we have concluded that there are real cases in which many kinds of applications could benefit from this resulting mainly in a cost optimization. This research can be extended by analyzing the results of dynamically allocated resources on multiprovider hybrid clouds in order to increase the cost savings.

Acknowledgements We gratefully acknowledge the work of our main Conference Manager tester, Abel Carril, in addition to our anonymous reviewers for their insightful and detailed comments. Development of Conference Manager was partially supported by the EU research project named Global Plaza.

References

1. Armbrust M, Fox A, Griffith R et al (2009) Above the clouds: a berkeley view of cloud computing. EECS Department, University of California, Berkeley
2. Barra E, Mendo A, Tapiador A et al (2011) Integral solution for web conferencing event management. In: IADIS int. conf. e-Society
3. Cerviño J, Escribano F, Rodríguez P et al (2011) Videoconference capacity leasing on hybrid clouds. In: IEEE CLOUD
4. Cerviño J, Rodríguez P, Trajkovska I et al (2011) Testing a Cloud provider network for hybrid P2P and Cloud streaming architectures. In: IEEE CLOUD
5. den Bossche RV, Vanmechelen K, Broeckhove J (2010) Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads. IEEE CLOUD, pp 228–235
6. Goyal P (2010) Enterprise usability of Cloud computing environments: issues and challenges. IEEE Enabling Technologies, pp 54–59
7. Hajjat M, Sun X, wei Eric Sung Y et al (2010) Cloudward bound: planning for beneficial migration of enterprise applications to the cloud. In: SIGCOMM
8. Leavitt N (2009) Is cloud computing really ready for prime time? Growth 27:5
9. Li A, Yang X, Kandula S et al (2010) CloudCmp: comparing public cloud providers. In: IMC. ACM, pp 1–14
10. Li H, Zhong L, Liu J, Li B, Xu K (2011) Cost-effective partial migration of VoD services to content Clouds. IEEE CLOUD, pp 203–210
11. Li J, Chinneck J, Woodside M et al (2009) Performance model driven QoS guarantees and optimization in clouds. In: ICSE workshop, CLOUD 2009, pp 15–22
12. Li J, Guo R, Zhang X (2010) Study on service-oriented Cloud conferencing. In: IEEE ICCSIT, vol 6, pp 21 –25
13. Song B, Hassan M, Huh EN et al (2009) A hybrid algorithm for partner selection in market oriented Cloud computing. In: MASS
14. Sotomayor B, Montero R, Llorente I et al (2009) Virtual infrastructure management in private and hybrid clouds. IEEE Internet Comput 13(5):14–22