Quality-Driven Optimal SLA Selection for Enterprise Cloud Communications

Pantelis A. Frangoudis*, Aggeliki Sgora‡, Martín Varela†, and Gerardo Rubino*
*INRIA Rennes-Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France
†VTT Technical Research Centre, Finland
‡Department of Informatics, University of Piraeus, 80, Karaoli and Dimitriou St., GR-18534, Piraeus, Greece
Email: *{pantelis.frangoudis, gerardo.rubino}@inria.fr, †martin.varela@vtt.fi, ‡asgora@unipi.gr

***Abstract*—With the availability of cloud computing infrastructures, migrating business-critical functionality to public clouds is becoming commonplace. Cloud providers typically offer a variety of computing capabilities and pricing options. Therefore, the problem of selecting the ones that suit enterprise needs becomes critical. Our major focus is on the migration of enterprise communication services, such as IP-telephony, to the cloud. We design a tool to assist in optimally deciding among the set of available hosting and network connectivity Service-Level Agreements (SLAs) under Quality-of-Experience (QoE) and budget constraints. In particular, we propose a multi-objective optimization framework making use of application-specific QoE estimation tools to tackle with the conflicting objectives of price and quality and demonstrate its application to a cloud-based teleconferencing service as a case study.**

I. Introduction

In order to reduce the maintenance cost and management overhead of dedicated infrastructures, and in some cases for the sake of improved reliability, enterprise communication services, such as messaging and multimedia conferencing are increasingly being migrated to public clouds [1]. The enterprise does not need to host dedicated servers and deal with issues of hardware and software redundancy and uptime, while at the same time it can save energy and space.

However, a new set of challenges emerges. Designing a smooth and disruption-free migration strategy, ensuring data confidentiality when moving critical functions outside enterprise premises, and maintaining or improving the level of service quality are typical concerns. The specific challenge we address is to optimally select among a set of available Service Level Agreements (SLAs) offered by cloud hosting and network service providers for connecting the enterprise sites with the cloud, facing a tradeoff between service quality and cost. Our notion of optimality therefore involves maximizing user quality while minimizing cost. We note that, often, the SLAs currently offered by some of the largest cloud providers are not at all suitable for this type of application.

Pantelis A. Frangoudis' and Aggeliki Sgora's work was carried out in part during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 246016. This work has also been supported by the CELTIC project QuEEN (Project ID: CP8-004), partially funded by Tekes, the Finnish Funding Agency for Technology and Innovation.

Our approach is *Quality-driven*: Taking into account the specific service-level objectives (SLOs), i.e., measurable performance parameters included in each SLA and the associated cost, and applying application-specific QoE (in particular, perceived quality) estimation tools, we propose an SLA selection framework which builds on a multi-objective optimization problem formulation. Our framework is suitable for cloud-based enterprise communication services, but has wider applicability. As a case study, we demonstrate its application to a cloud-based teleconferencing service, using quality estimation tools to predict user experience.

This paper is structured as follows. In Section II we present a review of relevant literature. In Section III we introduce our multi-objective optimization framework for Quality-driven SLA selection, which we apply to a cloud-based teleconferencing scenario in Section IV. We conclude the paper and discuss ongoing and future research directions in Section V.

II. Related work

A. QoE aspects on the cloud

QoE provisioning and management in a cloud setting is a challenging issue [2]. The fact that additional key actors mediate service provision adds complexity to both providing, but also monitoring the level of service quality offered to users.

Existing approaches focus on resource management within the cloud infrastructure and how it affects user experience. Kafetzakis *et al.* [3] study the problem of optimizing cloud resources based on QoE indications and propose a multidimensional architecture where agents residing within the cloud are used for monitoring QoE. Based on that, resource management and adaptation decisions are taken.

Our work bears similarities with the work of Qian *et al.* [4], in the sense that both focus on the QoE-based evaluation of cloud service providers. They propose a hierarchical model which builds on sub-models for cloud availability, output bandwidth, response time and latency and use the probability that service can be successfully provided (i.e., service is available and latency requirements are not violated) and the average service completion time as QoE indicators. They also study user request redirection strategies with the aim of maximizing the above QoE indicators. In our work, we use a QoE-driven evaluation of potential SLAs offered by

978-1-4799-4640-2/14/$31.00 ©2014 IEEE

cloud providers and apply quality estimators which are more suitable as indicators of QoE in the context of specific cloud communication services such as voice and multimedia.

To the best of our knowledge, using QoE as a tool for optimal SLA selection by a service provider is a topic not yet addressed.

B. Cloud networking

By design, in typical cloud architectures it is not always straightforward to manage the network. The network control plane is hidden and, in typical cloud SLAs, network guarantees are abstracted under generic availability and responsiveness provisions. For most cloud applications, in order for the service provider to accurately estimate and manage user QoE, it is critical to be able to measure and control network-oriented SLOs. The lack of control over network configuration challenges the migration of on-premise business and communication functionality to the cloud [5]. Real-time, delay-sensitive applications are more affected by this situation [6].

On the positive side, cloud providers have started offering direct connectivity options to their services. The aim is to establish private connectivity between the enterprise premises and the cloud, bypassing the public Internet. Amazon AWS Direct Connect [7], for instance, offers dedicated connections from corporate infrastructures (e.g., data centers or collocation environments) to the Amazon Web Service (AWS). Other providers also provide this type of service through third parties (e.g. dedicated connections from partner ISPs).

C. QoE assessment

In our work, estimating user experience is critical. QoE estimators are largely application-specific. In the area of voice communications, numerous approaches have been proposed [8]. Most suitable for our purpose are tools that can estimate subjective user experience using objective measurable parameters.

The E-model [9], is a parametric model designed to be used as a transmission planning tool, which provides an assessment of the combined effects of various transmission parameters in the mouth-to-ear path on user-perceived conversational voice quality. It takes into account a wide range of telephony-band impairments, in particular the impairment due to low bit-rate coding devices and one-way delay, as well as, the "classical" telephony impairments of loss, noise and echo. Based on the results from a large number of subjective tests done in the past on a wide range of transmission parameters, it can output a scalar quality rating value known as the "Rating Factor, R." R ratings can be transformed to Mean Opinion Scores (MOS), i.e. estimates of user opinion. While not optimal for accurate quality estimation, it provides a simple expression for quality, and it is often used in the literature as an estimator.

A more sophisticated and accurate objective technique is the Pseudo-Subjective Quality Assessment (PSQA) methodology [10], based on selecting a set of critical measurable parameters that affect QoE, performing subjective tests with human observers controlling these parameters and tracking user satisfaction, and training and validating a statistical learning tool with the test results. PSQA has been successfully applied for QoE assessment of listening quality for VoIP [11], interactive voice applications [12], P2P video streaming [13], and IPTV audio-visual assessment [14]. PSQA has also been applied as a tool for QoE-driven resource management and control [15], [16].

III. SLA selection framework

In this section we propose an optimization framework for QoE-driven SLA selection, which an enterprise can apply when migrating its communication services to the cloud. We formulate a multi-objective optimization problem and apply our modeling approach to a cloud teleconferencing scenario in Section IV.

A. Environment and assumptions

We consider an environment where the enterprise deploys a communication service (e.g., teleconferencing) on a public cloud, leasing a number of virtual machines (VMs) to deploy server instances. We will refer to the enterprise as the Service Provider (SP). In the context of a cloud teleconferencing scenario, each VM hosts a number of virtual conference rooms and we focus on a worst-case scenario, planning for the maximum number of expected simultaneous conference calls with the maximum number of participants in each call. Since the total number of virtual rooms is fixed, the number of conference calls to be hosted on each virtual machine (and, thus, the load of each VM) depends on the number of VMs leased. We assume that rooms are evenly distributed across VMs.

The SP needs to maximize the expected QoE of users (e.g., teleconference participants), and at the same time to minimize cost under budget and quality constraints. A set of SLAs are available for (i) cloud providers (CPs) and (ii) network service providers (NSPs), and the SP needs to select an optimal combination of SLAs and number of VMs to deploy. This problem naturally lends itself to a multi-objective optimization formulation.

For simplicity, but without loss of generality, we only consider the case of a single CP and a single NSP.

B. SLA specifications

The CP offers a set of infrastructure leasing options. The SP can lease a number of VMs with specific processing, storage and memory characteristics, where it can deploy its service and the accompanying SLA ensures a specific availability guarantees and the service credit that is returned in case of failure. In this work, we focus on the case where the service provider leases all the necessary infrastructure in advance and do not consider elasticity options, where the amount of cloud resources acquired is scaled with demand. The latter is a topic for future work.

At the same time, the SP sets up an SLA with a NSP to connect its premises with the cloud. We assume that the contract between the network provider and the enterprise will

provide specific bandwidth, delay and packet loss guarantees and will ensure specific credit returns in the event of SLA violation.

C. Problem formulation

We introduce the following notation:

- $S^{(H)}$: Set of available hosting (H) SLAs.
- $S^{(N)}$: Set of available network (N) SLAs.
- $p_i^{(H)}$: VM instance specifications for $S_i^{(H)}$, including CPU power, available memory and storage capacity.
- $p_k^{(N)} = \langle d_k, j_k, l_k \rangle$: (Worst-case, as per the proposed SLA) network connectivity characteristics for $S_k^{(N)}$, i.e. the delay (d_k), jitter (j_k), and packet loss ratio (l_k).
- $c_i^{(H)}$: Price per VM instance for $S_i^{(H)}$.
- $c_k^{(N)}$: Price for the overall network service for $S_k^{(N)}$.

The SP will deploy n VMs of a specific type to the cloud, selecting a combination $\left\langle S_i^{(H)}, S_k^{(N)}, n \right\rangle$, where $S_i^{(H)} = \left\langle p_i^{(H)}, c_i^{(H)} \right\rangle$ and $S_k^{(N)} = \left\langle p_k^{(N)}, c_k^{(N)} \right\rangle$.

Note that this model does not take into account service availability guarantees which are typically included in the SLAs, nor the respective credit returns. This is a topic we defer for future work.

The SP needs to attain two conflicting objectives: Maximize $Q(i,k,n)$, i.e., the minimum guaranteed QoE for its users while minimizing $C(i,k,n)$, i.e., VM hosting and connectivity costs when deploying n VM instances under SLAs $S_i^{(H)}$ and $S_k^{(N)}$. At the same time, it needs to respect minimum QoE (q_{min}) and budget (B) constraints. $Q(i,k,n)$ is application-specific; in Section IV we explain how such a function can be derived in a cloud-based teleconferencing context. We formulate this multi-objective optimization problem as follows:

$$\begin{aligned} &\underset{i,k,n}{\text{Maximize}} \quad F(i,k,n) = \begin{bmatrix} Q(i,k,n) \\ -C(i,k,n) \end{bmatrix} \\ &\text{subject to} \;\; Q(i,k,n) - q_{min} \geq 0, \;\; B - C(i,k,n) \geq 0 \end{aligned} \tag{1}$$

Since the two components of the objective function are conflicting, the above formulation expresses the tradeoff between quality and cost and, typically, there is no single solution which optimizes both. Therefore, our goal is to offer the service provider a set of attainable solutions, each representing a different preference as to the priority of each criterion.

We tackle this multi-objective optimization problem by applying a *scalarization* [17] approach: We instead solve a scalar optimization problem where we attempt to minimize the distance from an ideal (*utopic*) reference point. We select the weighted Chebyshev norm as our distance function, given by

$$L_\infty^{(F,w,z^u)} = \max(w_Q|Q(i,k,n) - Q^u|, w_C|C(i,k,n) - C^u|), \tag{2}$$

where $z^u = (Q^u, C^u)$ represents the utopic point in terms of quality and cost. The weight vector $w = \begin{bmatrix} w_Q \\ w_C \end{bmatrix}$ expresses the importance of each criterion in the selection of the final solution among a set of Pareto optimal[1] solution vectors for the original problem. It should be noted that by systematic variation of weights, minimizing the above function is able to reveal the Pareto optimal set even when the function to optimize is not convex [18].

An issue that needs to be tackled is that the two components of the objective function have different units and orders of magnitude (QoE vs. cost units). To this end, the functions need to be properly transformed via a normalization process. For a study and comparison of various function transformation schemes, see the work of Marler and Arora [19].

D. Solution complexity

An exact algorithm which exhaustively searches the parameter space to find the solution which minimizes the weighted Chebyshev distance from the reference point runs in $O(|S^{(H)}||S^{(N)}|B)$ time, where $|S^{(H)}|$ is the number of hosting SLAs, $|S^{(N)}|$ the number of network SLAs and B the budget constraint, further assuming that calculating $Q(i,k,n)$ takes $O(1)$ time. For each SLA combination, this algorithm calculates the attainable QoE and the solution cost for all possible numbers of deployed VMs such that the budget constraint is not violated[2]. The complexity of the algorithm is *pseudo-polynomial*: If we encode the input parameters (numbers of SLAs, SLA specifications, budget) in binary, then the input size of the algorithm is $O(log|S^{(H)}| + log|S^{(N)}| + (|S^{(H)}| + |S^{(N)}|)logP + logB) \approx O((|S^{(H)}| + |S^{(N)}|)logP + logB)$, where $logP$ is the maximum size of an encoded SLA specification in bits. Let $p = logP$ and $b = logB$; we then have that the input size is $O((|S^{(H)}| + |S^{(N)}|)p + b)$ and the time complexity is $O(|S^{(H)}||S^{(N)}|2^b)$. This is expected to scale poorly for large budgets. Note also that while computing the Pareto frontier, this algorithm is executed once for each different weight combination. However, in Section IV-D we show that this brute-force algorithm runs in acceptable time on a low-end workstation for realistic budget sizes.

IV. Use case: Cloud-based teleconferencing

A. Teleconferencing service

We assume a centralized teleconferencing system, where the server hosts virtual conference rooms. The conference server is a multipoint control unit (MCU) acting as a bridge for user streams. The MCU is responsible for decoding the incoming streams from all speakers in a room, mixing them and re-encoding them in a single stream for each participant. The more the participants and virtual rooms, the more the CPU overhead for the MCU due to mixing, and thus delay. For a detailed description of the architecture of a centralized teleconferencing service, the reader is referred to the work of Singh et al. [20].

[1] A solution vector for the original problem is Pareto optimal *iff* it is not possible to move from that point and improve at least one objective function without negatively affecting any other objective function.

[2] For a specific SLA combination, search stops after inspecting solutions with at most $\left\lfloor B / \min_i c_i^{(H)} \right\rfloor$ deployed VMs, since more VMs would result in a budget constraint violation.

B. Estimating perceived conversational quality

End-user quality of experience depends, among others, on a set of application-specific parameters, such as codec configuration, but also, importantly, on performance characteristics of the cloud hosting and network services. In our analysis, we consider application-related parameters such as audio bitrate, encoding algorithms, etc. as constant and focus on parameters which either can be directly expressed as SLOs (e.g., service availability, propagation delay, packet loss) or are a function of the SLOs and the characteristics of the leased infrastructure (e.g., delay introduced due to overloaded VMs).

The workload of each virtual machine has a direct impact on the latency experienced by end users. This adds up to network latency and affects user experience. The more the VMs leased by the SP, the less the load imposed on each of them and, thus, delay. For specific VM characteristics, a function of load vs. delay is necessary for QoE estimation.

To evaluate user experience, we apply automated, pseudo-subjective QoE estimation tools. Apart from various configuration parameters which we assume fixed, the input to these tools are end-to-end delay, jitter and packet loss rate and the output is a QoE estimate.

1) Cloud-induced delay: Given that the overall maximum number of conference calls is fixed and distributed evenly across the n deployed VM instances, the delay imposed to each conference call due to VM load is a function of VM specifications ($p_i^{(H)}$) and n. One option is to derive this function $d(p_i^{(H)}, n)$ empirically, by emulating simultaneous conference calls on a machine with these specifications and measuring the delay for VoIP stream mixing. Deriving such performance models is an interesting research topic in its own right. A relevant methodology for modeling application response times as a function of resource allocation and utilization in virtualized environments is proposed by Watson et al. [21]. They, as well as Bodík et al. [22], provide experimental evidence that response time increases linearly with load at low CPU utilizations, while non-linearities emerge beyond capacity and response times can become arbitrarily long. Therefore, a parametric model with a linear and an exponential term fits experimental data more accurately [22]. In these works, it was also observed that the curves of response time vs CPU utilization for different processor speeds look like linearly scaled versions of each other.

Starting from these observations, and to reduce measurement effort, but at the expense of accuracy, we propose to derive an empirical model of $d(p_0^{(H)}, n)$ based on experiments, where $p_0^{(H)}$ are the specifications of the lowest-power VM, and then approximate $d(p_i^{(H)}, n)$ for $i > 0$ by scaling $d(p_0^{(H)}, n)$ by a factor $b_i = \frac{CPU_0}{CPU_i}$, where CPU_i is the clock tick rate encoded in $p_i^{(H)}$. We illustrate our approach in the following example.

Assume a scenario with 3 available hosting SLAs, where the respective CPU speeds are CPU_0, $CPU_1 = 2CPU_0$ and $CPU_2 = 4CPU_0$. We aim to support a fixed number of teleconference calls n_c, with a fixed number of participants each. The n_c calls are evenly distributed across the n leased VMs such that each VM hosts $\frac{n_c}{n}$ calls. Let us further assume, as an illustrative example, that the expression of service delay in ms (response time) as a function of the number of calls x (load) hosted on a VM with CPU_0 power is given by $2x + e^{0.01x}$, which is an instance of the parametric model of Bodík et al. [22]. If we substitute x for $\frac{n_c}{n}$, we get an expression of delay as a function of the number of deployed VMs for the specific processor characteristics:

$$d(p_0^{(H)}, n) = \frac{2n_c}{n} + e^{0.01\frac{n_c}{n}}. \quad (3)$$

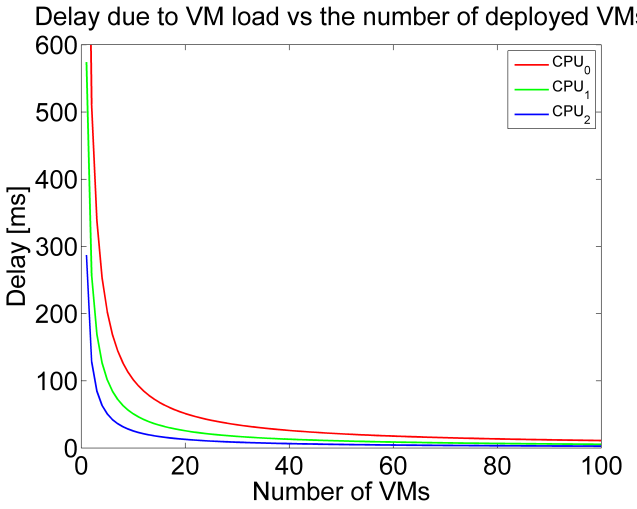

Fig. 1. Since the target total number of calls is constant and as the number of deployed VMs grows, the number of calls/VM (and thus, CPU load) decreases, which in turn reduces the delay experienced per user due to voip stream mixing. Each curve corresponds to VM types with different processing power.

Finally, we apply the scaling factors b_i to the above function to approximate $d(p_i^{(H)}, n)$, for $i > 0$. As an example, for $n_c = 500$, the effect of increasing the number of deployed VMs for each available hosting SLA is shown if Fig. 1.

2) QoE function: Despite the advantages of PSQA as presented in Section II-C, for the sake of simplicity and for saving space, here we apply the E-model for QoE estimation, since its closed-form expression shown in Eq. (4) is significantly simpler than that produced by PSQA (cf. [11] for more details). In particular, we utilize the methodology of Cole and Rosenbluth [23], where the E-model is reduced to directly measurable transport level metrics, selecting default values for all other parameters affecting conversational voice quality.

For a specific codec configuration (G.729a codec where each packet carries 20 ms of audio content, 60 ms jitter buffer at the receiver end) and assuming zero packet loss due to excessive jitter, the output of the E-model is given by the following formula:

$$\begin{aligned} R(d,l) = {} & 94.2 - 0.024 \cdot (d + 85) \\ & - 0.11 \cdot (d - 92.3) \cdot H(d - 92.3) - 11 \\ & - 40 \cdot \ln(1 + 10 \cdot l) \end{aligned} \quad (4)$$

where:

- d is the end-to-end network delay in ms
- l represents the percentage of packets lost in the network path
- $H(x) = 1 \quad \text{if} \quad x > 0; \quad 0 \quad \text{otherwise.}$

The R-factor can be mapped to a MOS using this formula [23]:

$$MOS(R) = \begin{cases} 1 & \text{if} \quad R < 0, \\ 4.5 & \text{if} \quad R > 100, \\ 1 + 0.035R + 7 \cdot 10^6 R(R-60)(100-R) & \\ & \text{if} \quad 0 < R < 100. \end{cases} \tag{5}$$

C. Objective functions

Regarding user experience, the additional processing delay due to stream mixing (see Section IV-B1) should also be taken into account in Eq. (4). Therefore, the objective function to maximize is

$$Q(i,k,n) = MOS(R(d(p_i^{(H)}, n) + d_k, l_k)). \tag{6}$$

At the same time, we need to minimize the cost function $C(i,k,n)$ to deploy n VMs under cloud hosting SLA $S_i^{(H)}$ and networking SLA $S_k^{(N)}$. This cost function is given by

$$C(i,k,n) = n \cdot c_i^{(H)} + c_k^{(N)}. \tag{7}$$

These functions have different units and orders of magnitude. We apply the *upper-lower-bound* approach [19] to normalize them, and the transformed functions follow:

$$Q^{(t)}(i,k,n) = \frac{Q(i,k,n) - Q^{min}}{Q^{max} - Q^{min}}, \tag{8}$$

$$-C^{(t)}(i,k,n) = \frac{C^{min} - C(i,k,n)}{C^{max} - C^{min}}, \tag{9}$$

where $Q^{(t)}$ and $-C^{(t)}$ are the transformed objective functions to maximize, and $*^{min}$, $*^{max}$ are the minimum and maximum values respectively that each objective function can take. Since Q is actually a MOS value in the 1-5 scale, $Q^{max} = 5$ and $Q^{min} = 1$. As to the cost objective function, this depends on the number n of VMs to deploy, which is one of the decision variables and unknown in advance. We can however limit its maximum value to our budget constraint B and its lower value to 0. Then, Eq. (8) and (9) become:

$$Q^{(t)}(i,k,n) = \frac{Q(i,k,n) - 1}{4}, \tag{10}$$

$$-C^{(t)}(i,k,n) = \frac{B - C(i,k,n)}{B}. \tag{11}$$

The transformed objective functions can now be used when minimizing (2). The problem to optimize thus becomes:

$$\begin{aligned} &\underset{i,k,n}{\text{Minimize}} \ \max(w_Q|Q^{(t)}(i,k,n) - Q^u|, w_C|C^{(t)}(i,k,n) - C^u|) \\ &\text{subject to} \ \ Q(i,k,n) - q_{min} \geq 0, \ B - C(i,k,n) \geq 0, \end{aligned} \tag{12}$$

TABLE I
CLOUD HOSTING SLA SPECIFICATIONS

Cost/VM ($/year)	CPU speed
358.64	1
717.68	2
1434.36	4
2859.96	8

where $w_Q + w_C = 1$ and $w_Q, w_C \geq 0$. As a reference point we use $C^u = 1$ and $Q^u = 1$, which are the optimal theoretical values for the transformed objective functions, representing the (unattainable) situation of a solution with maximum quality and zero cost.

The exact weight values are up to the SP's preferences and depend on how much the SP values the cloud telephony service. Intuitively, the more important the enterprise considers the cloud telephony service, the more it will be willing to spend to improve user QoE, always respecting the budget constraint. To visualize the tradeoff between quality and cost and offer the SP (decision maker) a complete view of the set of Pareto optimal solutions, we systematically vary the weights in (12) and we iteratively solve the optimization problem for each weight combination.

D. Numerical example

We demonstrate our methodology with a simple numerical example. The SP needs to select among a set of cloud hosting SLAs (Table I) and we assume the availability of network service SLAs (Table II) with varying delay and packet loss guarantees. CPU speed is expressed in normalized computation power units and we use actual prices from a large cloud provider for reserving a VM instance for constant operation for a one-year period. The maximum number of calls we wish to support is $n_c = 1000$ and the SP's budget is $B = 100000$. We solve the optimization problem by exhaustively exploring the solution space, calculating a Pareto-optimal solution for each weight combination. In Fig. 2 we show the set of Pareto-optimal solutions produced with the above algorithm. Each point represents a different tradeoff between quality and cost. The fact that we have only two objectives makes it easy to visualize the solution space, which facilitates decision making. In this example, for instance, it becomes evident to the SP that, after a specific point, increasing investment cost barely improves user QoE.

Performance-wise, calculating the Pareto-optimal points of Fig. 2 took approximately $0.9s$ on an Intel i3 workstation running Linux 2.6.35. Even for very large budgets (e.g., where $B/\min_i c_i^{(H)} \approx 30000$) and with 15 hosting SLAs and 15 network SLAs available, a single solution (for a single weight combination) takes less than $1s$ to compute.

V. CONCLUSION

We addressed the problem of selecting optimal combinations among available hosting and network connectivity SLAs when deploying a communication service to the cloud. We developed a multi-objective optimization framework capable

TABLE II
NETWORK CONNECTIVITY SLA SPECIFICATIONS

Cost ($/year)	Delay (ms)	Packet loss ratio
1000	200	0.01
2000	100	0.001
3000	50	0.0001
4000	15	0.00001
500000	12	0.000001

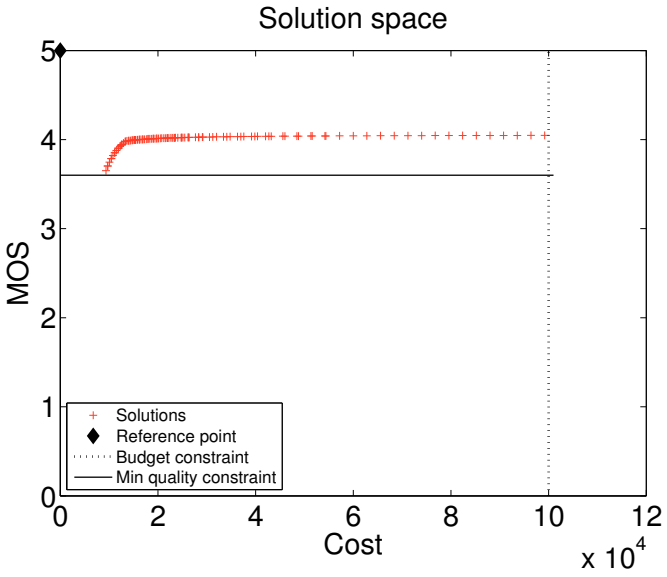

Fig. 2. The Pareto-optimal solutions shown in the figure were derived by solving the multi-objective optimization problem for multiple weight combinations. The straight lines represent the linear QoE and budget constraints.

of producing solutions which maximize the expected user experience while minimizing cost under budget and quality constraints. Since the two objectives are typically conflicting, our approach allows the decision maker to explore the solution space under varying weight combinations for them. Note that our problem formulation is generic: By applying the appropriate application-specific QoE methodology, it can be adapted to different use-case scenarios. As a case study, we have shown how our framework can be applied to a cloud teleconferencing service.

Our ongoing research is carried out on multiple fronts: We are working towards developing more sophisticated models of cloud response times, which are critical for accurate QoE estimation, while at the same time exploring alternative application scenarios for our framework. Finally, an important aspect for further research is to extend our work to consider elasticity options.

REFERENCES

[1] B. Haskings and A. Nilssen, "Migrating collaboration to the cloud," White Paper, Wainhouse Research, Jul. 2013.

[2] T. Hoßfeld, R. Schatz, M. Varela, and C. Timmerer, "Challenges of QoE management for cloud applications," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 28–36, 2012.

[3] E. Kafetzakis, H. Koumaras, M. Kourtis, and V. Koumaras, "QoE4CLOUD: A QoE-driven multidimensional framework for cloud environments," in *Proc. International Conference on Telecommunications and Multimedia (TEMU 2012)*, 2012, pp. 77–82.

[4] H. Qian, D. Medhi, and T. Trivedi, "A hierarchical model to evaluate quality of experience of online services hosted by cloud computing," in *Proc. IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2011, pp. 105–112.

[5] T. Benson, A. Akella, A. Shaikh, and S. Sahu, "CloudNaaS: a cloud networking platform for enterprise applications," in *Proc. 2^{nd} ACM Symposium on Cloud Computing (SOCC'11)*, Cascais, Portugal, 2011.

[6] K. Oberle, M. Stein, T. Voith, G. Gallizo, and R. Kubert, "The Network Aspect of Infrastructure-as-a-Service," in *Proc. 14^{th} International Conference on Intelligence in Next Generation Networks (ICIN)*, 2010.

[7] AWS Direct Connect. Amazon. [Online]. Available: http://aws.amazon.com/directconnect/

[8] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, "Quality of Experience of VoIP Service: A Survey of Assessment Approaches and Open Issues," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 491–513, 2012.

[9] ITU-T, "Recommendation G.107 - The E-model: A Computational Model for Use in Transmission Planning," 2011. [Online]. Available: http://www.itu.int/

[10] G. Rubino, "Quantifying the Quality of Audio and Video Transmissions over the Internet: The PSQA Approach," in *Communication Networks & Computer Systems*, J. A. Barria, Ed. Imperial College Press, 2005.

[11] Martín Varela, "Pseudo-Subjective Quality Assessment of Multimedia Streams and its Applications in Control," Ph.D. dissertation, INRIA/IRISA, Univ. Rennes I, Rennes, France, Nov. 2005.

[12] A. P. C. da Silva, M. Varela, E. de Souza e Silva, R. M. Leo, and G. Rubino, "Quality assessment of interactive voice applications," *Computer Networks*, vol. 52, no. 6, pp. 1179–1192, 2008.

[13] A. C. da Silva, P. Rodriguez-Bocca, and G. Rubino, "Optimal Quality-of-Experience design for a P2P Multi-Source video streaming," in *Proc. IEEE ICC*, May 2008, pp. 22–26.

[14] T. Mäki, D. Kukolj, D. Dordević, and M. Varela, "A Reduced-Reference Parametric Model for Audiovisual Quality of IPTV Services," in *Proc. QoMEX 2013*, Klagenfurt, Austria, Jul. 2013.

[15] M. Varela and J.-P. Laulajainen, "Terminal-Side QoE Estimations for Cross-Layer Network Control," in *Wired/Wireless Internet Communications*, ser. Lecture Notes in Computer Science, X. Masip-Bruin, D. Verchere, V. Tsaoussidis, and M. Yannuzzi, Eds. Springer Berlin Heidelberg, 2011, vol. 6649, pp. 140–149.

[16] J. Seppänen and M. Varela, "QoE-driven Network Management for Real-time Over-the-Top Multimedia Services," in *Proc. IEEE WCNC*, Shanghai, China, 7–10 Apr. 2013, pp. 1621–1626.

[17] K. Miettinen and M. M. Mäkelä, "On scalarizing functions in multiobjective optimization," *OR Spectrum*, vol. 24, no. 2, pp. 193–213, 2002.

[18] J. Koski and R. Silvennoinen, "Norm methods and partial weighting in multicriterion optimization of structures," *International Journal for Numerical Methods in Engineering*, vol. 24, no. 6, pp. 1101–1121, 1987.

[19] R. T. Marler and J. S. Arora, "Function-transformation methods for multi-objective optimization," *Engineering Optimization*, vol. 37, no. 6, pp. 551–570, 2005.

[20] K. Singh and H. Schulzrinne, "Centralized Conferencing using SIP," in *Proc. 2^{nd} IP Telephony Workshop (IPtel 2001)*, April 2001.

[21] B. J. Watson, M. Marwah, D. Gmach, Y. Chen, M. Arlitt, and Z. Wang, "Probabilistic Performance Modeling of Virtualized Resource Allocation," in *Proc. 7^{th} ACM International Conference on Autonomic Computing (ICAC '10)*, Washington, DC, 2010.

[22] P. Bodík, C. Sutton, A. Fox, D. Patterson, and M. Jordan, "Response-Time Modeling for Resource Allocation and Energy-Informed SLAs," in *Proc. Workshop on Statistical Learning Techniques for Solving Systems Problems (MLSys '07)*, Whistler, BC, Canada, 2007.

[23] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *ACM Computer Communication Review*, vol. 31, no. 2, pp. 9–24, 2001.