Soft Reservations

Uncertainty-aware Resource Reservations in IaaS Environments

Seyed Vahid Mohammadi[1], Samuel Kounev[2], Adrian Juan-Verdejo [1] and Bholanathsingh Surajbali[1]
[1]*CAS Software A.G, CAS-Weg 1-5, Karlsruhe, Germany*
[2]*Institute for Program Structures and Data Organization, Karlsruhe Institute of Technology, Karlsruhe, Germany*
vahid.mohammadi@cas.de, kounev@kit.edu, {adrian.juan, b.surajbali}@cas.de

Keywords: Cloud Computing, IaaS, Resource Reservations.

Abstract: Modern Infrastructure-as-a-Service (IaaS) provides flexible access to data center resources on demand in an elastic fashion to meet the highly variable workload requirements of cloud applications. Cloud providers aim to provision resources as efficiently and as quickly as possible to their consumers. However, the lack of information about the hosted applications and their workloads makes it hard for cloud providers to anticipate the future resource demands of their customers so that they can plan the capacity of their infrastructure. Cloud providers can receive arbitrary requests for allocating resources on-the-fly in a completely unpredictable manner. Given this unpredictability, it may happen that providers might not be able to provision the requested resources quickly enough, or in the worst case, they might ran out of capacity and may not be able to satisfy their customers resource demands. To address these concerns, in this paper we propose a new resource reservation mechanism, based on the concept of soft reservations, addressing the issue of uncertainty and lack of information concerning the expected future customer workloads and corresponding resource demands. The proposed resource reservation mechanism makes it possible for cloud providers to better plan the capacity of their infrastructure and continuously optimize the placement of virtual machines on physical nodes thus improving the infrastructure cost and energy efficiency. It also takes into account the uncertainty of resource demand estimations and enables proactive online capacity planning resulting in cost benefits for both cloud providers and cloud customers.

1 INTRODUCTION

Cloud Computing is emerging as a new computing paradigm providing cloud consumers (henceforth called consumers) with on-demand access to data center resources by integrating computing, storage and networking platforms in a transparent manner. One of the major factors for the success of cloud computing is its elasticity property and the pay-per-use pricing strategy.

Elasticity is one of the major essential properties of the cloud paradigm providing the ability to deal with load variations by automatically provisioning / deprovisioning resources on-the-fly to match the current demand, i.e., adding more resources during high load periods and consolidating the resources to fewer nodes when the load decreases.

Ideally, this implies that the amount of resources such as CPU, main memory and network bandwidth are assigned and utilized in an optimal manner. For a system deployed in a pay-as-you-go cloud environment, such as Infrastructure-as-a-Service (IaaS), elasticity is critical to minimize operating cost while ensuring acceptable performance during high load periods. It allows consolidation of the system to consume less resource and thus minimize the operating costs during periods of low load while allowing it to dynamically scale up as the load increases.

This elastic scaling is typically implemented using virtualization technology where consumers deploy their applications packaged in virtual machines (VMs) on a virtualized infrastructure. Each VM hosts a complete software stack (operating system, middleware, application components) and instances of the VM can be dynamically added or removed based on the load variation. This fine-grained allocation is referred in the literature as on-the-fly elasticity (Vijayakumar et al., 2010).

However, complex workload patterns highly affect elasticity. Indeed, time-varying workload intensities are already challenging to handle in todays Internet systems. Workloads can vary for multi-tier applications by orders of magnitude within the same business day which makes it hard for the cloud providers (referred as providers) to optimally allocate VMs. Through awareness of workload changes, providers can more effectively overcome this challenge with less difficulty in the future.

Consumers are in the best position to predict how their workloads would change over time. However, the separation of cloud providers (referred as providers) and consumers hinders the former in having access to such information. Providers do not have direct access to the applications running inside the hosted VMs. Therefore, they cannot predict the applications future workload needs and consequently their future resource demands. Similarly, consumers do not have access to the hardware infrastructure where VMs are deployed. Therefore, they cannot anticipate the effects of sharing resources with third-party applications deployed by the provider on the same virtualized infrastructure. Because of this lack of information, consumers can only specify the type and amount of resources (e.g., number of CPUs) that should be allocated to their VMs by means of resource reservations communicated to the provider.

Due to above described information gap, cloud providers are not in the position to predict the future resource demands of consumers so that they can plan the capacity of their infrastructure accordingly. Cloud providers can receive arbitrary requests for allocating resources on-the-fly in a completely unpredictable manner. Given this unpredictability, it may happen that providers might not be able to provision the requested resources quickly enough, or in the worst case, they might ran out of capacity and may not be able to satisfy the consumers resource demands. The latter may lead to violations of Service Level Agreements (SLA) leading to loss of customers and reputation for both the cloud providers and cloud consumers. As a result, in order to guarantee SLAs, cloud consumers are forced to reserve more resources than they actually need resulting in over-provisioning and associated over-subscribed costs.

We propose a new reservation mechanism in order to protect consumers and providers from the cost overhead incurred due to over-provisioning. In this reservation mechanism, consumers can issue pre-reservations, referred to as soft reservations and then claim the pre-reserved resources by issuing normal reservations, referred to as hard reservations, if they actually end up needing them. Soft reservations capture the estimated amount of resources that will be required by a consumer at a given future point of time as well as the probability of actually needing these resources. This approach also aims at closing the information gap between consumers and providers by supplying a communication mechanism to exchange the relevant information for both parties.

The proposed approach comprises mechanisms to exploit the exchanged information in a beneficial way for both consumers and providers. Consumers would be able to use low level information about utilization of physical resources to better estimate their actual resource demands for running their services at the desired Quality of Service (QoS) level. Meanwhile, providers would be able to exploit the information about the expected future resource demands of their consumers to better plan the capacity of their infrastructure and continuously optimize the mapping of logical to physical resources resulting in lower data center operating costs and energy consumption.

The rest of the paper is organized as follows: Section 2 describes our resource reservation approach. We describe types of information we need to exchange between consumers and providers in Section 3. In Section 4, we survey the previous works in this area. Our end-to-end envisioned approach is summarized in Section 5. Finally, we summarize the paper in Section 6.

2 RESOURCE RESERVATIONS

Cloud providers provide on-demand access to scalable computing, storage and networking resources over a wide-area network. Consumers are able to deploy the VMs required to satisfy their SLAs with their customers (SaaS end-consumers). Consumers may dynamically ask for resources by placing resource reservations with the provider to match their varying workloads and respective resource demands. Once a consumer submits a request to the provider the request will be accepted, provided that enough resources are available. Otherwise, it would be rejected or some other options could be offered, i.e., a counter-offer (Lu et al., 2011) may be made. If the request is accepted, the provider will need to find a mechanism to satisfy the request. Consumers take into account several aspects such as amount, level of granularity, validity period, certainty, and provisioning intervals when making reservations. For example, a reservation could look like:

> ”I need 10 nodes, each with 1 GB of memory, right now” or ”I need 4 nodes, each with 2 CPUs and 2GB of memory, from 2pm to 4pm tomorrow”.

The latter describes the amount of required resources for a specified time window (from 2 to 4 pm), whereas the former requests them immediately. Consumers

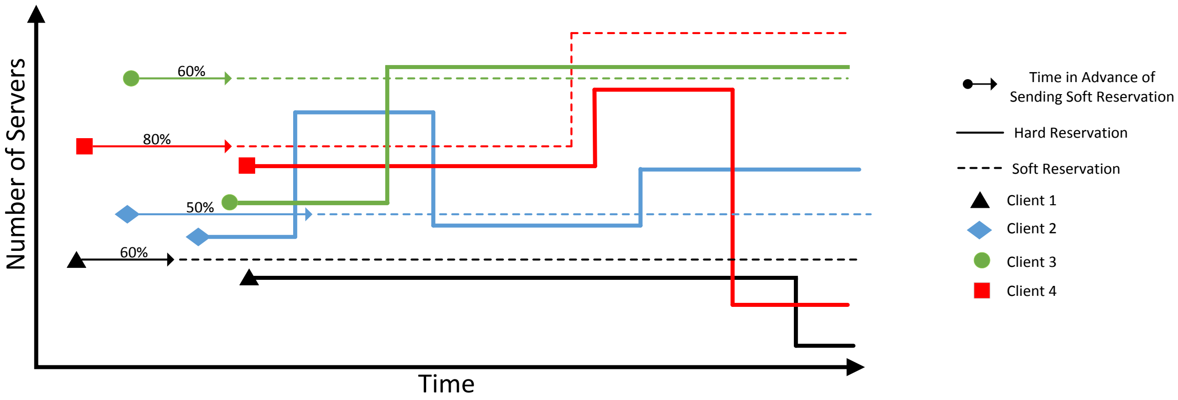

Figure 1: Snapshot of reservations received in infrastructure.

would normally not be able to predict in advance the exact amount of resources they will require due to the uncertainty about their future workloads and resource demands. However, consumers should normally be able to indicate how confident they are about the provided estimation. The latter can be done in different ways, e.g., by specifying the probability of actually requiring the reserved resources or by providing a cumulative distribution function of the amount of required resources.

2.1 Hard and Soft Reservations

As mentioned earlier, the actual resource utilization will only be known at runtime, so it is not possible for consumers to anticipate the real resource consumption in advance. In traditional resource reservation mechanisms, consumers have to pay for the requested resources even if they do not end up using them which is not aligned with the pay-per-use model in cloud computing.

Even though consumers may not be able to anticipate their exact resource demands, they would normally be able to approximately estimate the expected resource consumption based on workload forecasts and performance predictions with some level of certainty (Herbst et al., 2013). In the envisioned approach, consumers pre-reserve the forecasted required resources in long term through soft reservations. In these reservations, they basically specify how much resource they will need, in what time span, and how certain they are about their estimation.

The issuing of a soft reservation does not grant a consumer the requested resources. Nevertheless, once a consumer becomes more certain about his resource demands (near to the point of actual usage) traditional on-the-fly reservations (henceforth called hard reservations) can be issued to claim the resources that were previously pre-reserved through the corresponding soft reservations. If a consumer does not claim the required resources by means of hard reservations, the resources will not be allocated.

Figure 1 illustrates a snapshot of hard and soft reservations received by an IaaS provider and the time in advance of sending them by consumers over a period of time.

Consumers can choose only between using on-the-fly hard reservations or also taking advantage of the new soft reservations that allow them to book in advance resources from provider for a given time horizon (e.g., minutes, hours) in order to account for the perceived risk that a workload surge may occur. Informally, soft reservations will act as a form of insurance for consumers about obtaining resources at lower costs when needed provided that a correct bid for their future resource demands is communicated in advance to providers, whereas more expensive hard reservations (that are not preceded by a previous soft reservation) will be used to obtain unanticipated capacity that is required to process the current workload.

Soft reservations cater a win-win solution for both consumers and providers. For consumers, soft reservations will be much cheaper than hard reservations, since they only offer the right to obtain a set of future resources within a certain amount of time if they turn out to be required. If these resources are truly allocated at some point in the future, the consumer will have to pay additional compensation, but they will still save money compared to the on-the-fly hard reservations that would otherwise have to be used as workloads vary. If not allocated, the soft reservations would have instead merely served as an insurance policy for the consumer against high resource provisioning costs. However, the pricing model should provide a policy to avoid oversubscribed unclaimed soft reservations.

Similarly, providers will utilize the information provided through soft reservations as a basis for online capacity planning driving infrastructure management decisions. Upon observing changes in hard reservations, providers would dynamically allocate new capacity using standard mechanisms such as provisioning of new servers previously in stand-by mode. Providers can then use heuristics to optimize the placement for the new servers taking into account the currently active hard and soft reservations. Such heuristic algorithms should prioritize placements that improve the Total-Cost-of-Ownership (TCO) of the infrastructure.

Our proposed approach envisions different levels of soft reservations. Currently, four different dimensions are considered in order to define the softness level:

Provisioning interval of a soft reservation refers to the amount of time in which the softly-reserved resources are guaranteed to be provisioned if they end up being requested by issuing a respective hard reservation. The smaller the provisioning interval, the faster resources are guaranteed to be provisioned if they are claimed.

Validity period represents the validity time frame of a soft reservation. A reservation for one month would normally have higher importance and more implications for capacity planning than a reservation only for one day.

Time in advance represents the time in advance of sending a soft reservation before the desired period of its validity begins. A soft reservation for next week would normally have higher priority and more implications for capacity planning than a soft reservation with validity period beginning after one month.

Level of uncertainty refers to the estimated probability that the reserved resources will actually end up not being needed.

All four dimensions influence the degree of softness of soft reservations which in turn would normally influence the price for placing them. The softer reservations are, the cheaper their price would normally be expected to be.

Figure 2 illustrates the problems arises without soft reservations and benefits of the proposed hard and soft reservation compare to traditional mechanism. In this example, the provisioning interval and the level of uncertainty are fixed. The arrows depict the points in time at which soft reservation are submitted. One issue with the traditional reservation mechanism is the potential delay between the arrival of hard reservations and the actual provisioning of the requested resources (dashed red box on the left side of the first diagram). Another problem which might occur without soft reservations is that the provider might not be able to provision all of the requested resources (dashed red box on the right side of each diagram). The soft reservations help to address these two issues by enabling the provider to plan the capacity of the infrastructure such that all requested resources can be provided in time. The extent to which the softly reserved resources are guaranteed to be provisioned when placing a hard reservation depends on the four dimensions of the softness level explained above. Both exemplary problems described above are rooted in the inability of providers to anticipate what resources will be required by consumers in the future and thereby plan their capacity accordingly. The envisioned hard and soft reservation mechanism will tackle this problem by making it possible for consumers to communicate their estimated future resource demands.

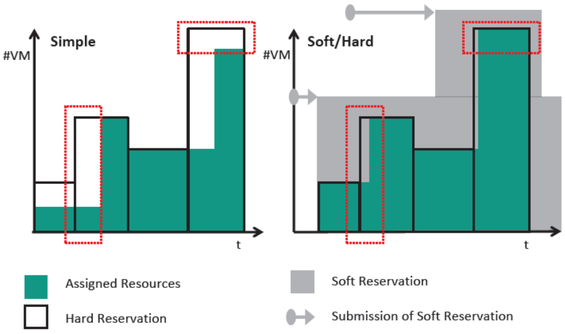

Figure 2: Example of hard and soft reservations.

3 INFORMATION EXCHANGE BETWEEN CLOUD PROVIDERS AND CONSUMERS

As mentioned earlier, the separation of cloud providers and consumers hinders providers to optimize their placement algorithms for provisioning IaaS resources. Moreover, the effects of sharing resources with third-party applications deployed in the same virtualized infrastructure are hidden from consumers. Hence, the exchange of information between the two stakeholders is beneficial for both. Providers can estimate the consumers future resource requirements and plan the allocation of VMs to Physical Machines (PM) improving the resource provisioning time and making significant cost savings at both ends. Moreover, information about the actual measured hardware utilization will allow consumers to only reserve and pay for the resources that are actually needed and to run their services at the service level required to fulfil the end-customers' (SLAs).

We identify two types of information which needs to be exchanged by consumers and providers, namely: (1) infrastructure monitoring-data supplied by the provider and (2) SLAs containing resource reservations placed by consumers.

3.1 Infrastructure Monitoring Data

Consumers suffer from the lack of direct access to the physical infrastructure level which is necessary to accurately monitor their resource consumption. Therefore, in order to enable consumers to accurately monitor and predict their future resource usage, providers must supply information and monitoring data about the physical infrastructure.

To provide such information, the provider needs to know the accuracy level and granularity of the measurement data required by each consumer. It is also important to know for how long historical data about resource utilization should be stored on the provider side in order to prevent infinitely growing large log files.

3.2 SLAs Containing Resource Reservations

Cloud providers normally do not have direct access to the applications and services running inside the VMs deployed on their infrastructure.

To bridge this gap, the proposed resource reservation mechanism offers a means for consumers to supply information about their expected future resource requirements based on workload forecasts and performance predictions. Such information should logically be part of the SLAs established between the consumers and providers.

The SLAs we are considering would not only cover classical metrics such as service response time and throughput, but also provide a powerful protocol for placing resource reservations, cancelling existing, or changing them. To realize these consumers need to know what types of resources are available for reservation and at what level of granularity they can be reserved (differentiating between general provider offerings, static agreements about the maximum allocations that could be provided to a consumer, and the availability of resources for possible reservation at a particular point in time).

4 STATE-OF-THE-ART

A vast amount of research exists in the literature on resource reservations in grid computing a summary of which can be found in (Rani et al., 2011). In cloud computing, advance reservations are an active area of research. In (Chaisiri et al., 2009), the authors present a stochastic integer program algorithm that works in an environment with multiple cloud providers. They propose an optimal virtual machine placement (OVMP) algorithm to minimize the total costs of reservations and on-demand resource provisioning. However, this approach does not consider any insurance policy for consumers allowing them to obtain their required resources at a cheaper price.

In (Mark et al., 2011), the authors address this problem in the same environment (Chaisiri et al., 2009), but they take different approach for handling future demands. They apply three different prediction algorithms (i.e., simple Kalman filter, double exponential smoothing, and Markov prediction) to predict the demands of customers, i.e., they use past usage history as a basis for forecasting future demands. In their approach, resource predictions takes place on the provider side while in ours, consumers are responsible for estimation of their future resource requirements.

Similarly, (Lu et al., 2011) provides a solution for the resource reservation problem in IaaS providers with limited resource capacity by which they are able to realize the feasibility of individual requests from consumers. If they are not able to satisfy the requests they will be able to provide an alternative offer by shifting requests in time (backward and forward) to fulfil them rather than refusing them. In their solution fragmentation in virtual resources is controlled and tried to be avoided. Resource requests are SLA-based and reservations take place during SLA negotiation. They utilize computational geometry for advanced reservation of resources.

Haizea [1] is a resource manager (”resource scheduler”) software component which allows consumers to request resources from a computational resource. Haizea uses leases as a basic resource provisioning abstraction. A lease is ”a negotiated and renegotiable agreement between a provider and a consumer, where the former agrees to make a set of resources available to the latter, based on a set of lease terms presented by the resource consumer”. In (Sotomayor et al., 2009), designers of Haizea, present a model for predicting various runtime overheads involved in using virtual machines, which efficiently support advance reservations. They extend Haizea to use this new model in its scheduling decisions, and use it with the OpenNebula virtual infrastructure manager so the scheduling decisions will be enacted in a Xen cluster.

In (Wang et al., 2011), the management of QoS

[1] http://haizea.cs.uchicago.edu

in the presence of resource reservations in cloud environments is investigated. In order to guarantee QoS in the near future and maximize the total revenue of the resource provider, resource reservation requests should be accepted selectively. The decision is made based on the analysis of the possible achieved QoS after resource configuration.

In (Diaz et al., 2011), three types of resource requests which are rejected because of finite number of (PM)s and to the variability of VM resources utilization is identified. (1) Immediate Rejection (IR): If there is not enough available capacity in any PM. (2) Resources Allocation Rejection (RAR): resources are allocated, and VMs are already hosted in PMs. However, due to the variability of VM workload, the sum of resources utilized by VMs hosted in the same PM can exceed its capacity. Therefore, one or more VMs must be suppressed to free the resources in PM. (3) Total of Rejections: It is the sum of IR and RAR ratios. They propose a new concept of Resource Over-Reservation (ROR) as a mean to reduce RARs. The basic idea is to pre-reserve additional resources in order to stick at the load variation. The authors found a trade-off between the IR and RAR as a value for ROR to keep percentages of total request rejections low.

To summarize, in all of these works traditional resource reservations were considered as an input to the algorithms for placement of VMs and none of them consider the gap of information between providers and consumers to address resource reservations in cloud IaaS environments.

5 APPROACH

Our contribution is new ”soft” resource reservation mechanism for consumers. As we described in Section 2.1, consumers can issue their soft reservations based on their expected resource demand in long term with some level of certainty. Typically this is close to the actual resource consumption point, when they become certain about their resource requirements, they can claim their softly reserved resources through hard reservations.

After receiving soft and hard reservations, providers can use them to continuously improve the quality of virtual machine placement decisions and to plan the capacity of their infrastructure. To be effective, soft reservations should be coupled with a pricing model that reduces risk and TCO for both providers and consumers (Rizou and Polyviou, 2012).

We also propose the importance of data sharing between consumers and providers which was described in Section 3. This sharing of information will help both parties to characterize and dynamically react to unexpected changes of usage patterns of services and systems.

Figure 3 summarize our end to end envisioned approach. Consumers automatically adapt the amount of resources requested from providers based on the dynamic performance models which will be automatically calibrated at run-time. These predicted resources will be translated to soft and hard reservation and will be sent to providers. Providers adapt the mapping of requested logical resources to physical resources in a dynamic way and strictly accounting for TCO. One possibility to realize the online forecasting and performance predicting process is using the Descartes Meta Model (Huber et al., 2012). Within this approach, consumers and providers will individually work towards achieving better performance and TCO, thus promoting a more efficient use of data center resources at reduced cost. Particularly, consumers will automatically adapt the amount of resources requested from providers based on dynamic performance and TCO models which will be automatically maintained, calibrated, composed and evaluated at runtime (Kounev et al., 2011). This will produce continuously TCO optimization.

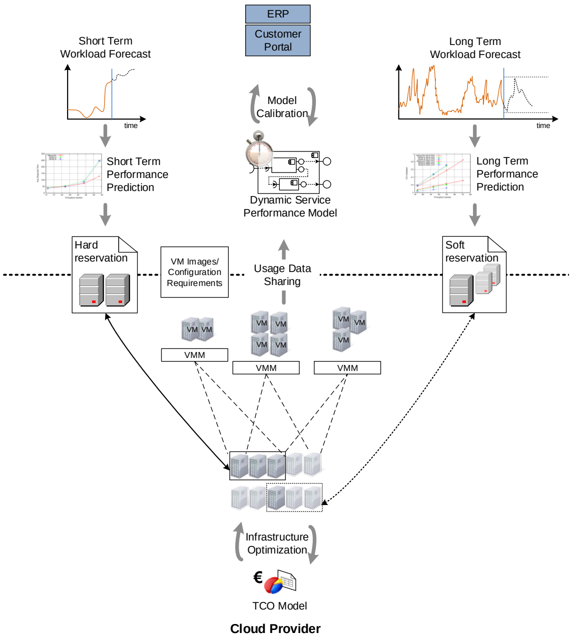
ERP
Customer Portal
Short Term Workload Forecast
time
Long Term Workload Forecast
time
Model Calibration
Short Term Performance Prediction
Dynamic Service Performance Model
Long Term Performance Prediction
Hard reservation
VM Images/ Configuration Requirements
Usage Data Sharing
Soft reservation
VM VM
VMM
VM VM VM VM
VMM
VM VM VM
VMM
Infrastructure Optimization
€
TCO Model
Cloud Provider

Figure 3: Online system performance model.

6 CONCLUSIONS

In this paper, we introduced our research roadmap through the definition of soft and hard reservations.

This mechanism enables consumers to communicate their workload forecasts in long term to providers by means of soft reservations. Consumers can claim the softly reserved resources when they become certain about their estimated resource requirements. Soft reservations act as an insurance policy that guarantees that the consumer will receive the softly reserved resources with cheaper price once they are requested through hard reservations. Similarly, providers have to send hardware utilization data (with respect to privacy of other customers) to the consumer. By means of these types of reservations, consumers would reserve and pay for the amount of resources they use and will receive their requested resources faster. Providers would be able to estimate the amount of resources they should provide in any point in time. Therefore, they would be able to manage their resources more efficiently by keeping PMs off or turn off PMs. In our future work, we intend to develop algorithms on the provider side to handle these reservations. These algorithms will cater for determining the expected required capacity at a given point of time in the future. Furthermore, our algorithms will identify future changes in capacity needs and will optimize the on-the-fly placement of VMs taking into account the costs of different reconfiguration options.

ACKNOWLEDGEMENTS

The research presented in this paper has been supported by the European Union within the FP7 Marie Curie Initial Training Network "RELATE" [2].

REFERENCES

Chaisiri, S., Lee, B.-S., and Niyato, D. (2009). Optimal virtual machine placement across multiple cloud providers. In *Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific*, pages 103 –110.

Diaz, F., Doumith, E., and Gagnaire, M. (2011). Impact of resource over-reservation (ror) and dropping policies on cloud resource allocation. In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, pages 470 –476.

Herbst, N. R., Huber, N., Kounev, S., and Amrehn, E. (2013). Self-Adaptive Workload Classification and Forecasting for Proactive Resource Provisioning. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering (ICPE 2013), Prague, Czech Republic, April 21–24*.

Huber, N., Brosig, F., and Kounev, S. (2012). Modeling Dynamic Virtualized Resource Landscapes. In *Proceedings of the 8th ACM SIGSOFT International Conference on the Quality of Software Architectures (QoSA 2012), June 25–28, 2012, Bertinoro, Italy*. Acceptance Rate (Full Paper): 25.6%.

Kounev, S., Brosig, F., and Huber, N. (2011). Self-Aware QoS Management in Virtualized Infrastructures (Poster Paper). In *8th International Conference on Autonomic Computing (ICAC 2011), June 14–18, 2011*, Karlsruhe, Germany.

Lu, K., Roblitz, T., Yahyapour, R., Yaqub, E., and Kotsokalis, C. (2011). Qos-aware sla-based advanced reservation of infrastructure as a service. In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, pages 288 –295.

Mark, C., Niyato, D., and Chen-Khong, T. (2011). Evolutionary optimal virtual machine placement and demand forecaster for cloud computing. In *Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on*, pages 348 –355.

Rani, B., Venkatesan, R., and Ramalakshmi, R. (2011). Resource reservation in grid computing environments: Design issues. In *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, volume 4, pages 66 –70.

Rizou, S. and Polyviou, A. (2012). Towards value-based resource provisioning in the Cloud. In *2012 IEEE 4th International Conference on Cloud Computing Technology and Science*.

Sotomayor, B., Montero, R., Llorente, I., and Foster, I. (2009). Resource leasing and the art of suspending virtual machines. In *High Performance Computing and Communications, 2009. HPCC '09. 11th IEEE International Conference on*, pages 59 –68.

Vijayakumar, S., Zhu, Q., and Agrawal, G. (2010). Dynamic resource provisioning for data streaming applications in a cloud environment. In *Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science*, CLOUDCOM '10, pages 441–448. IEEE Computer Society.

Wang, X., Xue, Y., Fan, L., Wang, R., and Du, Z. (2011). Research on adaptive qos-aware resource reservation management in cloud service environments. In *Services Computing Conference (APSCC), 2011 IEEE Asia-Pacific*, pages 147 –152.

[2] http://www.relate-itn.eu/