# Network-aware Cloud Brokerage for telecommunication services

Giuseppe Carella, Thomas Magedanz
Technische Universität Berlin
Berlin, Germany
giuseppe.a.carella@tu-berlin.de, tm@cs.tu-berlin.de

Konrad Campowsky, Florian Schreiner
Fraunhofer FOKUS
Berlin, Germany
konrad.campowsky@fokus.fraunhofer.de,
florian.schreiner@fokus.fraunhofer.de

***Abstract*—Cross-domain Cloud Brokering mechanisms enable elastic and cost-efficient utilization of cloud resources distributed across multiple cloud platforms. They are allowing cloud service providers to cost-efficiently exploit the growing competition in the cloud provider market.**

**Existing elastic cloud computing solutions are optimizing cloud resource utilization merely within a specific cloud provider platform. Optimized multi-site Cloud Brokering mechanisms on the other hand enable economically efficient cloud resource consumption. However, in order to satisfy QoS requirements of cloud-based, real-time multimedia telecommunication services, enhanced QoS assurance mechanisms for multi-site cloud brokers need to be in place.**

**In this paper we optimize and evaluate the FOKUS Cloud Broker solution through experimentation in a multi-site cloud testing facility which allows experimentation under different networking conditions, i.e. over standard, best-effort internet connections across several cloud platforms in Europe, as well as under fully controllable network conditions.**

**The result of this work shows the benefits of network-aware cloud brokering mechanisms. Moreover, this paper shows the terms under which additional real-time data on network performance is useful for enhancing cloud brokering mechanisms, especially for meeting QoS requirements of real-time communication services. This work also shows how initial, service-specific correlation of network, service and host performance parameters, furthermore improves the overall cloud brokering performance.**

***Cloud brokerage, QoS, network performance, multi-domain, cross-platform, multi-cloud, Future Internet, Internet of Services***

## I. INTRODUCTION

Cloud computing mechanisms have already gained broad attention attracting steadily increasing numbers of service providers by providing means to optimize resource consumption and means allowing for outsourcing of infrastructure and service management costs as well as by enabling pay-as-you-go cost models.

Elastic cloud computing, defined as the capability of cloud platforms to dynamically up- and down-scale resources according to current demand, is one of the most important mechanisms of a cloud platform, especially of an Infrastructure as a Service (IaaS) cloud platform, as it allows efficient cloud resource utilization.

By utilizing converged, all-IP, access-network-independent service control platforms such as the IP Multimedia Subsystems (IMS) [1] an increasing number of telecommunication operators and service providers are currently consolidating their service infrastructures towards converged Next Generation Network (NGNs) service delivery platforms (SDPs). Although these SDPs are sought to greatly reduce new telecommunication service's time-to-market, based on re-usable service enablers, significant up-front service infrastructures investments as well as significant operational expenditures are usually still required. With cloud computing mechanisms applied to IMS-based service infrastructures, IMS service providers are charged on a pay-per-use basis, significantly lowering the risk of unsuccessful investments. However, whereas cloud-based Web-services are already widespread, telecommunication service providers, having significantly higher QoS requirements, predominantly are still reluctant to move their services to external clouds. This is because in most cases cloud platforms are either not QoS-aware at all, or unable to assure end-to-end service qualities. Only after QoS levels can be assured, by also taking into account the network performance between telecommunication core network and cloud-based service platform, more telecommunication service providers will be willing to move their value-added services to the cloud.

By providing the required flexibility to dynamically choose amongst the currently best cloud platforms in terms of QoS, but also in terms of costs, cloud brokering mechanisms are providing important benefits to service providers. Service providers want to dynamically select a cloud platform for hosting their services, which provides optimal QoS levels at the best price. Therefore cloud brokering mechanisms need to find the optimal trade-off between current costs and QoS levels, based on user preferences. Here different preferences need to be supported, ranging from highly cost-sensitive and QoS insensitive preferences (e.g. for best-effort service providers), to highly QoS-sensitive and cost-insensitive preferences (e.g. service providers providing guaranteed QoS levels to premium customers). This work is aiming to provide a well-balanced, customizable solution, trimmed to and optimized for the QoS requirements of a specific service.

978-1-4673-2798-5/12/$31.00 ©2012 IEEE

This work is mainly motivated by the following rationale: firstly, current elastic cloud computing mechanisms (such as Amazon's Elastic Compute Cloud [2], Rackspace, CloudSigma and ElasticHosts) per se do not support dynamic and seamless migration of services between multiple cloud provider infrastructures and platforms, thus fostering cloud provider lock-ins rather than empowering service providers to exploit the increasing competition in the cloud provider market. Other solutions in fact do support brokerage across several different cloud platforms, such as RightScale [3], however none of these offerings are sensitive to network performance (between core network and cloud service platforms). This usually affects typical Web services to a lesser degree, but to a much higher degree affects real-time communication service's quality (voice / video, conferencing, messaging,) as these services are significantly more sensitive to the actual network performance.

Secondly, IMS-/NGN-based telecommunication service platforms can indeed be deployed on multiple cloud platforms. Even a cloud-based IMS core platform offering IMS as a service is currently investigated and no more a far-out vision. However, as telecommunication services providers usually require guaranteed QoS levels, well-balanced QoS-aware and cost-aware cloud resource placement strategies are required. Telecommunication service providers would never utilize even the cheapest cloud platform, if the provided QoS would not satisfy their customer's minimum QoS requirements. And vice versa, even if the delivered QoS of a particular cloud provider platform outbids any competing cloud offering, if customers are satisfied with the provided QoS, service providers would select the cheaper solution.

The presented work investigates mechanisms for optimizing cloud platform selection and elastic brokerage under QoS and cost constraints. The FOKUS Cloud Broker Engine (CBE), for telecommunication services, as shown in Figure 1. is capable of simultaneously interworking with multiple cloud platforms via standard cloud computing interfaces. Furthermore, the CBE is capable of dynamically up- and down-scaling of cloud resources and able to dynamically migrate cloud resources across multiple cloud platforms.

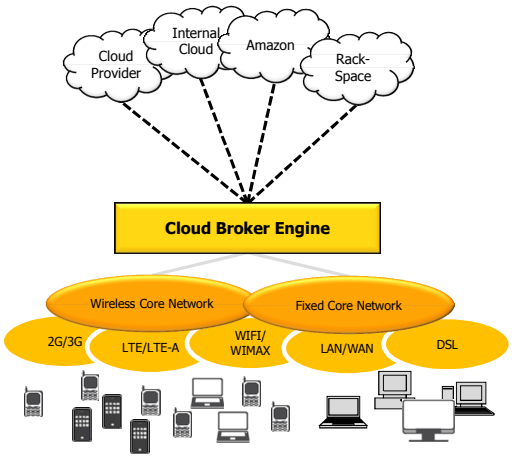


Figure 1. Cloud Broker Engine

The design and the components of the CBE have already been introduced in [4], where also the basic algorithms for elastic up- and down-scaling of cloud resources across multiple cloud platforms are described. In [5] we benchmarked the CBE's resource allocation performance and service migration performance in a local testbed setup under controlled network conditions. In this paper we evaluate the newly developed QoS vs. Cost optimization algorithm of the CBE in a real-world scenario, deploying the CBE on top of the large-scale, multi-site cloud infrastructure of the EU FP7 project BonFIRE [6]. A real-world IMS VoIP service is elastically deployed across different cloud platforms. After having initially benchmarked the VoIP quality of the service under different network conditions (packet loss and jitter), the CBE optimization algorithm dynamically selects the optimal cloud platform in terms of costs (simulated) and QoS during a typical day according to pre-configured user preferences, where QoS as well as cost constraints are initially specified by CBE users. It shows that indeed network performance degradations as well as cost variations should both be taken into account in order to optimize cloud platform selection and cloud resource allocation.

The remainder of this paper is structured as follows. Section II provides the necessary background information as well as a related research section. Section III roughly describes the functional design of the CBE solution, while Section IV describes the actual CBE optimization algorithm. Section V describes the validation of the CBE in BonFIRE's multi-site cloud infrastructure. Finally section VI concludes the paper providing an outlook on next steps.

## II. BACKGROUND AND RELATED WORK

Several works in the field of cloud computing have already been presented, but only very limited part of them focused on cloud-based IMS / telecommunication infrastructures and services.

Whereas the authors of [7] focus on a specific service - a cloud-based IMS presence service - without considering QoS parameters, the authors of [8] are focusing on Web Services having a similar approach but with a limited number of analyzed monitored data and using non-weighted round robin load balancing algorithm. In [9] a "profile-based" solution is being described, which only takes into account the CPU utilization of a given Virtual Machine (VM).

In [10] the possibility to deploy two different cluster-based services on top of a virtualized infrastructure is being analyzed. Authors are using, similarly to our solution, a hybrid cloud infrastructure, but they do not consider real-time monitoring data to scale their services automatically.

Authors of [11] are considering cloud brokering algorithms for optimizing VMs placement, but they are not considering monitoring of deployed services qualities to provide guaranteed QoS levels. Nevertheless, the approach in [11] is designed in an extendable way, and could easily take into account network and service quality performance parameters for optimizing the cloud platform selection process.

## III. CLOUD BROKER ENGINE ARCHITECTURE

The cloud broker solution, already described in [4] and [5], basically consists of a distributed monitoring system (local monitoring agents deployed in VMs and load balancers LBs, and a central monitoring aggregator), a rules engine (where thresholds and basic rules are stored and evaluated), a generic API for interoperating with different private and public clouds (OpenNebula [12], OpenStack [13], amazon EC2 [2], and the core intelligence - the CBE.

Based on real-time monitoring data (network performance, node/VM performance and utilization, service performance, service load) the CBE, after querying the rules engine, controls single or multiple cloud management systems (located in multiple, distributed cloud platforms) for optimizing cloud resource deployment, through elastic scaling and migration mechanisms and thereby optimizing service quality.

### A. CBE Optimizaion algorithm

In order to dynamically select the optimal provider in terms of QoS and costs at each given point in time, we utilize an algorithm, which, based on user preferences, takes into account not only the static values introduced by the user for a provider, but also real-time monitoring information such as the service execution time (measured at the load-balancing component), as well as network performance measurements (actively and passively measureable; we use Iperf for active jitter and loss measurements) in an easily extensible way.

The first step of this algorithm involves the user who specifies his preferred Key Performance Indicators (KPI) (i.e. cost, service execution time, QoS parameters like VoIP quality) and their weight in relation to each other. We define by $K_j$ the KPIs for a provider, with $\lambda_j$ the specific weight assigned by the user for a specific KPI, and with m being the number of KPIs.

$$\sum_{1}^{m} \lambda_j = 1$$

After this step, by utilizing the parameters introduced within a deployment file, the system is able to dynamically select the optimal provider at each given point in time.

In order to normalize different KPIs and to create ranking table, we define with $V_{ij}$ the value for a KPI $K_j$. For those KPIs for which the values are better if lower $(KPI-)$ we define:

$$NV_{ij} = \frac{max\ - V_{ij}}{max} \lambda_j\ (for\ KPI-)$$

Where $max$ is defined as the maximum acceptable value for these KPIs. For those KPIs for which the values are better if greater $(KPI+)$ we define:

$$NV_{ij} = \frac{V_{ij} - \min}{V_{ij}} \lambda_j\ (for\ KPI+)$$

Where $min$ is defined as the minimum acceptable value desirable for these KPIs.

By doing so, a new table is created mapping Providers to KPIs with normalized values. The choice of the best provider is to determine the provider with highest values for those KPIs where the best value is the higher one and with lower values for those KPIs where the best value is the lower one. Defining with KPI+ the first one, and with KPI- the last one, we define:

$$T_i = \sum_{j=1}^{m} NV_{ij}$$

Where $T_i$ is the score of the i-th provider. The best provider at any given time is the one with the highest value of $T_i$.

## IV. EVALUATION OF CBE OPTIMIZATION

One of the main assertions of this work is that without knowing the correlation between resource utilization and service quality, no fully optimized solution can be found. Therefore, for each new service to be deployed on a multi-site cloud, we benchmark the service quality against the systems' utilization and against important network performance parameters. For the IMS-based VoIP announcement service, used in the evaluation, we selected the Perceptual Evaluation of Speech Quality (PESQ) parameter (i.e. ITU-T standard for end-to-end speech quality assessment [13]) to be the most important parameter for defining the actual service quality.

### A. Service Benchmarking and CBE Calibration

By generating an increasing number of requests, we calibrated the CBE by determining the CPU threshold (here, as shown in Figure 2. at a 77% CPU utilization level) above which the PESQ significantly deteriorates. Determining this threshold is already important for optimizing resource consumption (i.e. number of VMs, costs, energy). Generic (non-application-specific) approaches either lead to overprovisioning of resources (i.e. unnecessary costs, energy consumption) or under-provisioning (i.e. likelihood of QoS degradation).

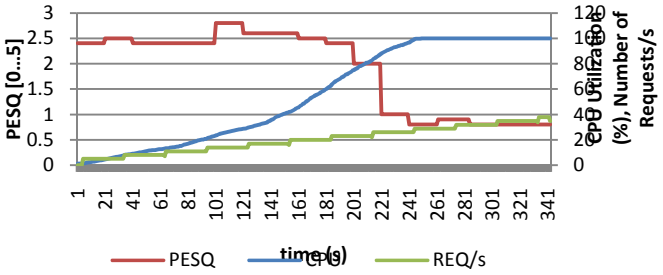


Figure 2. Calibration PESQ vs. CPU

Having done that, we are also interested in determining the impact of network performance on the service quality. As bandwidth limitations with our narrowband VoIP service are unlikely to impact the voice quality, we focused on packet loss and jitter. Figure 3. shows the impact of packet loss on the voice quality of our selected SEMS media application. It shows that only after a significant packet loss (here ~25% packet loss), PESQ values fall below 2 (i.e. "poor" voice quality).

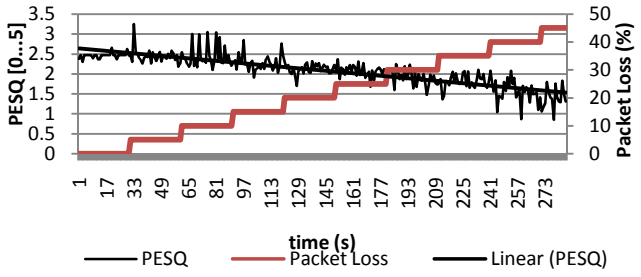


Figure 3. Calibration PESQ vs. Packet Loss

Another important network parameter affecting VoIP quality is Jitter, the variation of delay between incoming packets. We realized that PESQ values can already deteriorate with only Jitter values (as shown in Figure 4. 30-40 ms Jitter). This is pretty much in line with the ITU-T recommendation on "Network performance objectives for IP-based services" [15], where acceptable Jitter values for VoIP services are defined to be below 50ms.

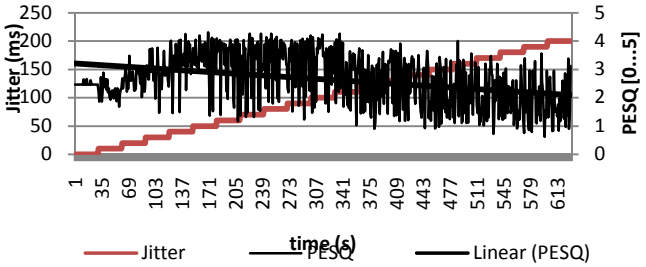


Figure 4. Calibration PESQ vs. Jitter

### B. *Large-scale, multi-cloud Testbed Setup*

In order to evaluate the CBE in a real world, multi-cloud scenario, we utilized pan-European BonFIRE [6] infrastructure, a unique testing facility for cloud-based services and systems. BonFIRE is comprised of multiple cloud sites, located in different European countries, such as the HLSR cloud platform in Germany, the INRIA cloud platform in France and the EPCC cloud platform in the UK as utilized in this tesbed setup, shown in Figure 5.

The Open IMS Core [16], a reference implementation of the 3GPP IMS specification (including Proxy-, Interrogating- and Serving Call State Control Functions P-/I-/S-CSCF and the Home Subscriber Server HSS) as well as the SIP Load Balancer are deployed on the German cloud platform from HLRS, whereas the French (INRIA) and the UK (EPCC) cloud platform are hosting the actual media server [17] instances.

We utilize IMS Bench SIPp [18], an SIP/IMS load generator that conforms to the European Telecommunications Standards Institute (ETSI) IMS/NGN Performance benchmark specification [20], for generating constant load as well as load variations. We use a modified version of the Kamailio load balancing software [19] that supports weighted round-robin SIP load balancing (an algorithm efficiently utilizing serving nodes in the back-end). We utilized a RTP-proxy integrated in the P-CSCF node to avoid NAT traversal problems.

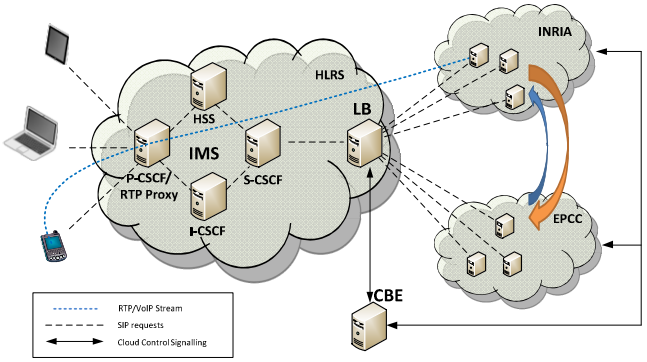


Figure 5. Test Setup

The utilized SIP Express Media Server (SEMS) [17] offers different services like announcement and conferencing. For the experiment with duration of 1050 minutes, i.e. 17 hours and 30 minutes, we dynamically loaded, scaled and migrated the SEMS VoIP announcement service, a standard telephony, e.g. announcement on non-available callee.

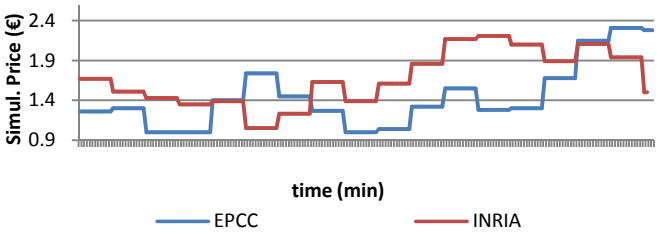


Figure 6. Simulated Price of Cloud Resources / Platform

Since BonFIRE's multi-site cloud is a non-commercial platform for experimentation, we simulated varying cloud resource prices during the experiment, as shown in Figure 6. On the one hand, intra-day cloud resource price variations are already a commodity, exploited by commercial offerings like SpotCloud [21], on the other hand, in order to make heterogeneous cloud resource costs fully comparable, initial cloud resource performance benchmarking is required, as developed and studied in [22]. Only after application specific cloud resource benchmarking has been conducted, the "true" costs (e.g. the provided computational performance of a cloud computing resource per unit of price) can be determined. Being only peripherally in the scope of this work, we utilized identical VM images, presuming that the simulated "price" in a real world scenario would a-priori take into account comparable performance / price benchmarks for each specific cloud sites' resources (e.g. micro, small, medium, large VM instances on Amazon).

During the experiment, we measured jitter between the IMS site and each cloud site hosting the media/announcement service. As shown in Figure 7. jitter between the IMS cloud and each media service cloud kept at a low comparable level for most of the time.

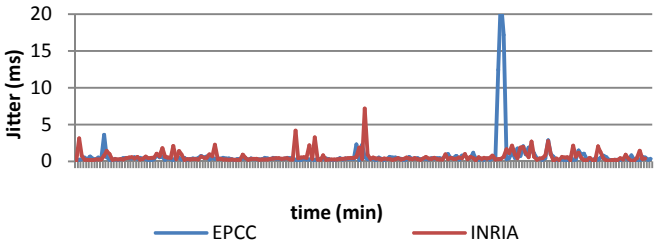


Figure 7. Jitter measurements

Only for a period of approximately 30 minutes we measured a significant increase of Jitter of up to 25 ms between the IMS cloud at HLRS and the EPCC cloud.

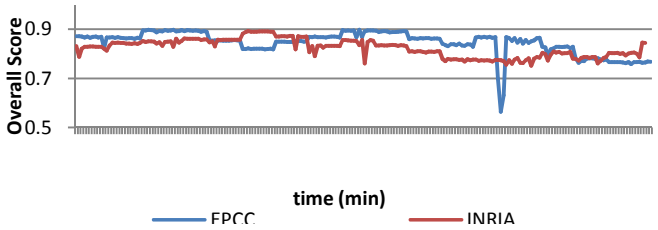


Figure 8. Calculated Score

Both, cost variations as well as the service quality variation (here the Jitter influence on PESQ) determine the score of each particular cloud platform at each particular point in time as shown in Figure 8.

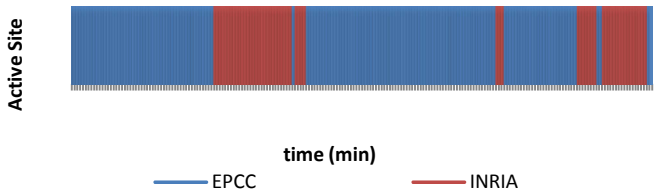


Figure 9. Dynamically selected Cloud Platform

Based on the current score, the CBE selects the currently best cloud platform for each particular cloud service. In our experiment, the selection of the optimal cloud platform was mainly dominated by the current simulated price of each platform. Only, as shown in Figure 9. , during the aforementioned 30 minutes of significant increase of jitter, the cloud site selection process was determined by QoS related observations. As measured initially, during the calibration phase, in our case jitter of more than 30ms already deteriorated the VoIP quality / PESQ to a level below 2, i.e. “poor” in some cases. Therefore, the QoS related decision to switch cloud sites was appropriate, although 25ms jitter in this experiment-run is still a low value, compared to other days where we measured up to 160ms Jitter with a duration of up to one hour on the same link.

Not shown here, but also relevant to mention, based on the above described PESQ versus CPU utilization calibration, and based on CBE benchmarking mechanisms previously described in [5], the CBE is also able to efficiently cope with moderately varying load situations (load variations which tolerate a delay of elastic deployment of additional resources of up to 45s) without PESQ degradation. Optimization mechanisms for the CBE’s elastic resource utilization performance are a topic on their own and have been described previously in [4], and will be further optimized based further testing on BonFIRE.

## V. Conclusion and Future Work

Being able to dynamically utilize cloud resources across different cloud platforms provides a number of advantages for telecommunication service providers. On the other hand by being able to flexibly operate service environments in flexible hybrid manner telecommunication service providers are able to cost-efficiently utilize internal and external cloud resources, greatly reducing the risk of unsuccessful infrastructure investments, but still being able to assure acceptable QoS levels. Enhancing QoS-awareness of cloud brokering mechanisms is not only important for optimized resource utilization (as resources are neither over- nor under-provisioned), but also mandatory for telecommunication services providers in order to be able to assure and guarantee service quality levels.

We believe that only by analyzing the impact of user load, network performance and cloud resource utilization parameters on a specific services’ quality, an optimal, resource and cost efficient strategy for elastic scaling of cloud resources, as well as cloud site selection (including migration) can be found.

While this work, at the current stage, does not claim to provide a fully optimized solution, an indicative strategy for optimizing elastic cloud resource utilization mechanisms across multiple cloud sites is introduced. The shown feasibility study and the evaluation of the developed cloud broker engine on top of a real-world multi-site cloud facility for large-scale experimentation confirms applicability and reasonableness of the proposed approach. Indeed optimization of QoS versus cost for cloud-based services is important and beneficial.

We are currently investigating enhanced mechanisms for the cloud site selection algorithm, where we try to quantify the anticipated benefit (e.g. better QoS, lower costs) of switching between two or more sites. Only if the benefit is high enough we would want to select an alternative cloud platform, reducing the likelihood of repeatedly occurring, sporadic cloud site switching. This will avoid ineffective switching, especially because cloud resources can usually not be purchased on a per-second/per-minute basis. Furthermore we investigate load and cloud performance prediction mechanisms for additional trimming of the optimizer’s performance. In order to proof the commercial utilization of the overall system, we are currently conducting tests on Amazon EC2 later also Rackspace, CloudSigma and ElasticHosts, where we foresee to be able to prove the importance of QoS awareness, especially for telecommunication services.

## Acknowledgment

This work has been partially funded by EU FP7 Integrated Project BonFIRE [6]. The BonFIRE project has received research funding from the EC's Seventh Framework Programs (EU ICT-2009-257386 IP under the Information and Communication Technologies Program).

## REFERENCES


[1] 3GPP. TS 23.228. IP Multimedia Subsystem (IMS) .

[2] Amazon. Amazon web services. 2011; Available from: http://aws.amazon.com

[3] RightScale. Cloud Management for public and private clouds, http://www.rightscale.com

[4] P. Bellavista, K. Campowsky , G. Carella, L. Foschini, T. Magedanz, F. Schreiner, "QoS-aware elastic cloud brokering for IMS infrastructures", The Seventeenth IEEE Symposium on Computers and Communications (ISCC'12). July 1 - 4, 2012, Cappadocia, Turkey.

[5] K. Campowsky , G. Carella, T. Magedanz, F. Schreiner, "Optimization of Elastic Cloud Brokerage Mechanisms for Future Telecommunication Service Environments", Praxis der Informationsverarbeitung und Kommunikation. Volume 0, Issue 0, ISSN (Online) 1865-8342, ISSN (Print) 0930-5157, DOI: 10.1515/pik-2012-0036, June 2012

[6] EU FP7 BonFIRE Project: http://www.bonfire-project.eu

[7] T. Louvain-la-Neuve, "Migration of Mobicents Sip Servlets on a cloud platform." Master Thesis, Universite catholique de Louvain, Louvain, School of Engineering, academic year 2010 – 2011, online: http://thibault.leruitte.name/tmp/cc04980d.pdf

[8] W. Iqbal, M. Dailey, and D. Carrera, "Sla-driven adaptive resource management for web applications on a heterogeneous compute cloud," in Cloud Computing, ser. Lecture Notes in Computer Science, M. Jaatun, G. Zhao, and C. Rong, Eds. Springer Berlin / Heidelberg, 2009, vol. 5931, pp. 243–253.

[9] Y. Jie, Q. Jie, and L. Ying, "A profile-based approach to justin-time scalability for cloud applications," sep. 2009, pp. 9–16

[10] R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, "Elastic management of cluster-based services in the cloud," in ACDC '09: Proceedings of the 1st workshop on Automated control for datacenters and clouds. New York, NY, USA: ACM, 2009, pp. 19–24.

[11] J. Tordsson, R. S. Montero, R. Moreno-Vozmediano, I. M. Llorente, Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers, Future Generation Computer Systems 2012

[12] The OpenNebula Project: http://www.opennebula.org

[13] Open source software for building private and public clouds: http://www.openstack.org/

[14] ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, online: http://www.itu.int/rec/T-REC-P.862/en

[15] ITU-T Y.1541, "Network performance objectives for IP-based services" International Telecommunications Union, Geneva, Switzerland (12/2011).

[16] The Open Source IMS Core, http://www.open-ims.org

[17] The SIP Express Media Server: http://www.iptel.org/sems

[18] IMS Bench SIPp, Open Source IMS benchmarking tool, http://www.sipp.sourceforge.net/ims_bench

[19] Kamailio the Open Source SIP Server: http://www.kamailio.org/w/

[20] European Telecommunications Standards Institute. Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS/NGN Performance Benchmark. ETSI TS 186 008-1, October 2007.

[21] SpotCloud, Cloud Capacity Clearing House, Spot Market, http://spotcloud.com

[22] Phillips, S.C., Engen, V., and Papay, J. "Snow White Clouds and the Seven Dwarfs." in Proceedings of the IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom). 2011.