QoE4CLOUD: A QoE-driven Multidimensional Framework for Cloud Environments

Emmanouil Kafetzakis*, Harilaos Koumaras†, Michail Alexandros Kourtis‡, and Vaios Koumaras§
Institute of Informatics & Telecommunications, National Center for Scientific Research “Demokritos”,
P.O. Box 60228, GR-15310 Agia Paraskevi, Greece
Email: {*mkafetz, †koumaras, ‡akis.kourtis, §vaios}@iit.demokritos.gr

***Abstract*—Cloud Operators, in order to respond effectively to the QoS requirements of cloud applications, are obliged to apply over-provisioning policies. In general, this tactic leads to severe waste of the available cloud computing resources. Similarly, both Service/Platform Providers and End-Users wish to avoid the extra cost of this over-provisioning tactic and pay only per use, without having to statically reserve extra resources in advance. As a consequence, the cloud community urgently asks for flexible and intelligent management solutions, towards enhanced and efficient utilization of the cloud. The ultimate target of any intelligent cloud management scheme should be the provision of a service at an adequate quality level, creating the need to introduce the notion of Quality of Experience (QoE). In this context, this paper proposes a unified QoE-aware management framework, directly targeting to cloud computing environments. The proposed management system suggests the optimization of cloud resources usage and offered services in terms of QoE, satisfying the different service and resource requirements of all the involved cloud entities. In addition, the proposed novel approach merges together various QoS aspects in a multidimensional framework, referred to as QoE4CLOUD, which considers the perceived quality as the key metric for the management and performance optimization of the cloud environment.**

***Index Terms*—Cloud, Quality of Experience, Quality of Service, Resource Management**

I. INTRODUCTION

During the recent years, cloud computing architectures have become popular as a method of deployment of workloads and for delivering of computing applications as a service rather than a product. The concept of clouds has demonstrated commercial success, while it is expected to attain an even larger part over the next decade (see, e.g., [1], [2], [3]). The potential of clouds is leveraged by the fact that it allows the reduction of entry cost for new services. Therefore, cloud computing minimizes the business establishment cost, the investments on new infrastructure and optimally lowers the risk of launching a new service, platform or product.

Cloud computing represents a model for enabling ubiquitous and on-demand network access to a shared pool of configurable, computing resources (e.g., networks, servers, storage, applications, services, etc.) that can be rapidly provisioned and released with reduced management effort and service provider interaction [4]. In this concept, cloud computing mainly provides three different types of services: Software as a Service (SaaS) (e.g., SalesForce [5], Basecamp [6], GoogleApps [7], etc.), Platform as a Service (PaaS) (e.g., Windows Azure [8], Force.com [9], Google App Engine [10], etc.), and Infrastructure as a Service (IaaS) (e.g., Amazon AWS [11], Rackspace.com [12], Cloud Hosting [13], etc.). None of the three types of services requires end-user knowledge of the physical location and the configuration of the cloud that delivers the services.

In such a composite computing ecosystem, three different entities are getting involved: the Cloud Operator (i.e., the owner/provider of cloud resources), the Service/Platform Provider (i.e., the user of the Cloud Operator’s infrastructure) and the End-User (i.e., the final user that consumes services provided by the Service/Platform Provider). Each one of these entities competes with the others for optimizing its own Quality (since the resources of a cloud are finite and specific). For example, the issue of Quality of Service (QoS) is a critical factor for the success of a cloud Service/Platform Provider, considering that if the requested service is not delivered as expected, it may tarnish the provider’s reputation, diminish the revenues and finally devastate the business model. To make things worse, a legacy Service Level Agreement (SLA) (i.e., the legally binding contract stating the QoS guarantees required by the Service/Platform Provider or End-User) cannot directly be applied in cloud computing environments. This is because the typically included functional parameters (e.g., maximum response time, throughput factor, error rate, etc.) and non-functional parameters (e.g., timeliness, scalability, availability, etc.) have context-dependent meaning and thus difficult to manage. Furthermore, all these parameters have different quality impact for all the involved cloud entities.

Apart from assessing the capability of clouds in order to guarantee a certain level of performance (i.e., provision of QoS guarantees), it is valuable to understand cloud entities expectations, experience, and overall satisfaction. In other words, it is necessary to develop a framework focused on the actually perceived quality (i.e., Quality of Experience - QoE) received by the named cloud entities. Such a framework should also consider business perspectives (e.g., the cost in order to achieve a specific perceptual level, etc.), which are included in the term Quality of Business (QoBiz). As a consequence, the management of emerging cloud platforms is more likely to become a multidimensional problem, taking into account QoE, QoS and business-related issues (i.e., QoBiz).

We now outline the remainder of this paper. Section II briefly reviews some relevant works, while Section III presents the challenges and defines the QoE-driven cloud management

978-1-4673-2781-7/12/$31.00 ©2012 IEEE

concept. In the next section, the multidimensional approach for QoE is further considered. Finally, we conclude in Section V.

II. Recent Advances in Cloud Computing

A cloud computing environment must be elastically scalable; in other words it must have the ability to flexibly expand as the offered load and the business demands change. However, this feature requires the development of a diverse set of algorithms, similar to those outlined below. The study of Elastic Scalability and QoE Assessment for Cloud Services are prerequisites for the construction of an intelligent QoE Management and Control mechanism for the cloud resources. In the rest of this section, we describe the state-of-the-art in these three technological aspects of the cloud ecosystem.

A. Advances in Elastic Scalability

The first step for succeeding elastic scalability in cloud environments is the resource discovery and the proper monitoring of the cloud resources. Traditional monitoring technologies for single machines or clusters are restricted to locality and homogeneity of monitored objects and therefore, cannot be applied in the cloud in an appropriate manner [14]. There are lots of third-party collectors of cloud statistics which provide monitoring facilities (e.g., Cloudkick [15], Nimsoft Monitor [16], Monitis [17], Opnet [18], RevealCloud [19], etc.). All of them are proprietary solutions and do not aim at defining a standard for cloud monitoring.

The next step towards elastic scalability is the modelling of computing resources and the definition of QoE requirements of different cloud entities. In the case of cloud environments, it is crucial that the resource modelling is able to represent virtual resources, virtual networks, and virtual applications. Therefore, the existing service architectures should be expanded to include the virtual resources, described in terms of properties and functionalities [20]. The level of abstraction that will be selected must consider as many details as possible, and at the same time permit that the problem of resource allocation/ optimization to be tractable.

B. Advances in QoE Assessment for Cloud Services

Before being able to offer services/platforms in a cloud ecosystem without degradation in QoE, work should be done on the definition of QoE metrics. As already mentioned, these metrics are different for each entity of the cloud. Traditionally, the service performance described in an SLA includes response time, utilization, throughput, delay, jitter, availability, etc. These metrics may suffice for the description of the service offered from the Cloud Operator to the Service/Platform Provider. However, the QoE that the End-User understands results from the combination of the Internet QoS and the QoS delivered from the Cloud Operator to the Service/Platform Provider. Therefore, it is highly desirable for Service/Platform Providers to be able to evaluate the QoE received by End-Users by a single metric. This metric is expected to include the QoS offered by the Cloud Operator, the Service/Platform Provider and the Internet's QoS. Reference [21] firstly identifies interactions among the cloud entities and afterwards evaluates the QoE for the End-Users in this complicated environment. Similarly, [22] evaluates Amazons Grid Computing services, while [23] presents a methodology for selecting Cloud Operator. Cloud computing hardware reliability issues are studied in [24], [25]. Work [26] also examines host reliability issues, but from the perspective of the End-Users. To the best of our knowledge, there do not exist other works that follow similar multidimensional QoE approach as this proposed here.

This paper contributes to the evolution of existing QoE metrics towards multidimensional assessment and measures of cloud system quality, as perceived and reported by the End-User at the levels of PaaS and SaaS. Multidimensional approaches are required to holistically evaluate the End-User QoE using a single metric. Currently, few scientific studies into QoE-relevant factors exist for cloud service ecosystems [27], [28]. Beyond scientific results, companies like Infovista [29] or Compuware [30] offer proprietary, closed assessment solutions for monitoring quality at an IaaS level.

C. Advances in QoE Management & Control

The construction of an intelligent cloud management system raises a number of challenging problems. For example, the resource manager must possess precise and updated information on the resource usage at any particular time and place within the cloud. Any change in the resource usage may cause an activation event in order to make real-time reallocation of available resources for the fulfillment of QoE requirements. To this aim, fast and effective workload balancing solutions may be employed [31]. Before describing the contributions of the paper to the management and control of cloud ecosystem, some relevant works are presented.

Publication [32] proposes a middleware architecture for enabling SLA-driven clustering of QoS-aware application servers. After applying load balancing techniques, the resource usage of application servers is optimized, and the application hosting SLA is fulfilled without creating resource overprovisioning costs.

References [33], [34], [35], [36], [37] provide stable service levels for applications and platforms being hosted in a cloud ecosystem, as described in their SLA agreements. In particular, [33] proposes a three-step approach to map SLA and QoS requirements of business processes to cloud infrastructures. When a performance gap occurs, translucent replication of services is employed. From another point of view, [34] prioritizes Video-on-Demand (VoD) traffic considering the respective charging model. The total revenue of the service provider is maximized, through the definition of a proper optimization problem. The pay-as-you-go billing approach adopted in Cloud computing challenges resource provisioning for service providers. In this context, [35] based on the Dirichlet multinomial model presents an efficient reputation-based QoS provisioning scheme. The cost of computing resources is minimized, while the desired QoS metrics are satisfied. Publication [36] applies feedback control theory to present a Virtual Machine (VM)-based architecture for adaptive management of virtualized resources in cloud computing. Moreover, it models an adaptive controller that dynamically adjusts multiple virtualized resources utilization to achieve application SLA in cloud computing. Unlike the previous counterparts, [37] considers also power management issues, since it studies a service cloud environment with mobile devices.

As it was emphasized, one of the most important benefits of cloud ecosystem is its ability to allow the use of on demand resources. This feature leads to dynamically scalable systems and platforms. To take full advantage of the benefits of dynamic scaling, a cloud client (user or middleware) needs to be able to make accurate decisions on when to scale up and down. The scaling decisions must be done in advance to compensate for the overhead of using virtual resources, specifically their setup time. Therefore, a prediction method is usually used, as it best suits this task. Article [38] presents a new approach to the auto-scaling problem based on identifying past patterns that are similar to the present use of the system.

III. Challenges and Solution Concepts

The current cloud platforms (e.g., VMware vSphere [39], Microsoft Hyper-V [40], Real Hat [41]) virtualize the host's physical resources (through the use of specific software called hypervisor) and make them available to multiple guest virtual machines. These virtual machines share the various cloud platforms resources concurrently. The most of the cutting-edge cloud management solutions (e.g., vCloud Director [42], Abiquo [43], DynamicOps [44], Gale Technologies [45], Platform Computing [46]) create resource pools and automatically allocate resources among virtual machines based on pre-defined rules and policies. The resources (e.g., processor usage, memory usage, storage, etc.) are managed at the cloud level rather than the machine-by-machine level. Also, all the resource allocation policies proposed so far are static and can be executed only at the initialization phase of the virtual machines. Finally, all available solutions are proprietary, and there are no widely accepted standards or de facto open source reference implementations for cloud management applications.

As expected, in order to respond effectively to the QoS requirements of cloud applications, over-provisioning policies are widely followed. However, both the Service/Platform Provider and the End-User want to pay-per-use, without having to statically allocate resources in advance. Therefore, the cloud community urgently asks for flexible management solutions, tailored to the existing and future needs of the Service/Platform Providers and the End-Users.

Towards this goal, we propose a unified QoE-aware management framework, directly targeting cloud computing environments. The proposed management framework suggests the optimization of cloud resources and services in terms of QoE, satisfying the different service requirements of all the involved cloud entities. It should be emphasized that the management framework introduce a multidimensional approach to QoE that supersedes the legacy one (see Section IV for the proposed QoE framework description).

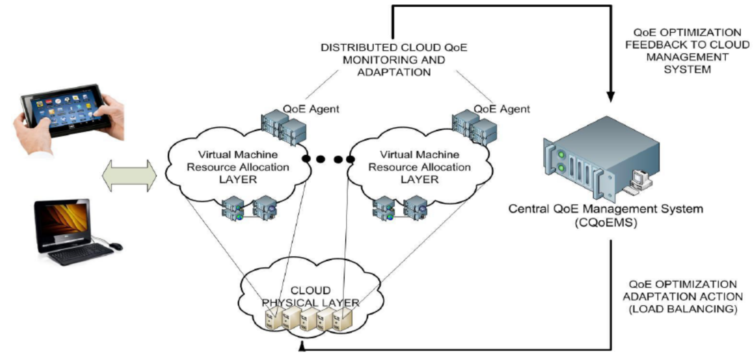

Fig. 1. QoE monitoring and provision scheme for the cloud computing environment.

This paper deals with QoE-driven resource optimization and proceeds beyond the one-dimensional QoS/QoE approach by considering QoS, QoE and QoBiz, and including all cloud entities. The highly composite environment of cloud computing is approached through a layered abstraction model, according to which, a specific module manages the IaaS layer towards optimizing QoE at the PaaS and at the SaaS layer. The IaaS layer is selected for enforcing the management decisions since it is the most flexible. It provides control over the hypervisor, the operating systems, the storage, and the network interfaces. In contrast, the PaaS and SaaS layers give control only to the hosting environment configurations and to the End-User specific application configuration settings, respectively.

According to the paper's concept, both the QoS characteristics at the PaaS layer and the QoE offered to the SaaS layer form the overall cloud platform QoE that is the basis for the quality-aware cloud management decisions. It is worth noting that the proposed QoE-driven management framework is market-oriented, since the notion of QoE as used here is considered in terms of End-User QoE as the currency, by including QoBiz.

Towards the overall QoE optimization, the management framework has to constantly monitor and evaluate the QoE condition of each service hosted in the virtual machines inside the cloud execution platform. The QoE assessment and the related local parameters should be reported from distributed QoE Agents (both at service initialization and at run-time stage) to the Central QoE Management System (CQoEMS) (see Fig. 1).

The CQoEMS module may provide sophisticated data mining and analysis methods for deriving the measures used for the overall QoE optimization of the cloud services. As most cloud tools, it should organize physical nodes into clusters [47] and place workloads on the right cluster in a timely and cost-effective manner. Following load balancing techniques, it should distribute workloads across two or more cloud clusters in order to optimize resource utilization and maximize the overall QoE.

Moreover, CQoEMS is responsible for the provision of

QoE guarantees to the services of high priority. The expected level of service between the Cloud Operator and the Service/Platform Provider or the Service/Platform Provider and the End-User can be described in proper SLAs after negotiations between the involved cloud entities. As an innovation for cloud services, the SLAs may be defined in terms of End-User QoE. However, QoE changes constantly and needs to be closely monitored by the QoE Agents.

The QoE Agents are in charge of reporting the QoE assessment of each service hosted in the virtual machines of the cloud. This report helps the resource reconfiguration to be based on the overall assessment of the QoE demands and the available resources. In other words, the QoE Agents dynamically provide a QoE optimization feedback to the cloud management system, aiming at the optimization of the delivered QoE and the satisfaction of the respective SLAs.

The QoE Agents are proposed to be placed within the cloud computing infrastructure in order to eliminate possible false quality degradation detections. More specifically, in case of the placement of a QoE Agent at the End-User, the detected quality degradation caused by an external non-cloud parameter (e.g., due to impairments in the access network between the End-User and the cloud) could have been falsely attributed to the cloud infrastructure/allocated resources. Since it is not possible for the QoE-Agent located at the End-User's side to differentiate the cause of the reported quality degradation, this in turn, will lead to faulty load balancing decisions from the CQoEMS.

Therefore, we propose that the most efficient placement for the QoE Agent is within the cloud infrastructure and not at the End-User side. In this way, the QoE Agents are able to:

- Assess the QoE that is expected to be delivered to the End-User.
- Detect quality degradations that are caused purely due to internal cloud computing parameters (e.g., shortage of available resources, etc.).
- Eliminate the false reported quality degradations that are caused due to external factors, usually in the access network between the Cloud and the End-User.

Therefore, following our architecture, the QoE Agents are able to monitor, assess and accurately report the QoE degradations caused by internal factors of the Cloud to the CQoEMS, avoiding false incidents by external/non-cloud parameters.

In the case that the expected QoE strongly differs from the actual QoE delivered to the End-User, the management scheme has to resort to a combination of probes placed within the Cloud and at the End-User side. This combined placement allows for the proper attribution of external non-cloud problems to QoE-impairments. This choice also reflects approaches followed, e.g., by multimedia network tomography/fault isolation methods, where assessment is performed at different service-chain locations to enable proper fault isolation (see, e.g., [48], [49]). However, this combination of probes inserts extra communication overhead (i.e., the End-Users QoE measurements exchange to cloud-based QoE Agents or directly to the CQoEMS), and have to be followed only if it is necessary.

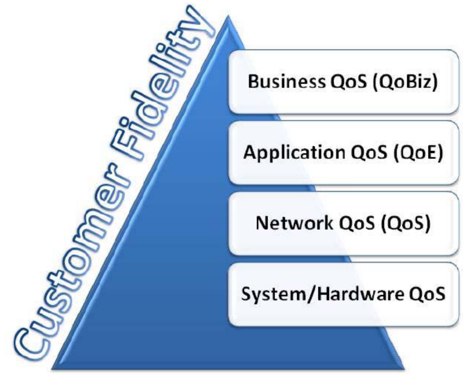

Fig. 2. The novel QoE-driven multidimensional framework (QoE4CLOUD).

IV. The Novel Multidimensional Approach for QoE

Beyond existing, proprietary solutions for monitoring IaaS or PaaS cloud services (see, e.g., [29], [30], [50]), we introduce a novel approach for linking different aspects of quality: the system/hardware-related QoS aspects, the network QoS aspects, the QoE aspects and the business-related aspects (QoBiz). As a result, our approach merges all these QoS aspects into a multidimensional framework referred to as QoE4CLOUD, which considers the QoE as the key metric for driving the quality assessment and the optimization procedures. Therefore, one of the major contributions of the paper is the introduction of QoE4CLOUD as the framework that manages quality with unified criteria in cloud computing ecosystems[1].

The proposed QoE framework approach enables the analysis of the aforementioned four different QoS aspects and as result, it provides a feedback-based convergence process. Ultimately, the convergence process conduces to the overall QoE optimization, required to guarantee the customer's fidelity and the Cloud Operator/Service Provider's profitability (see Fig. 2).

In the lower layer, the QoE4CLOUD approach considers and analyzes all the issues that the Cloud Operator should take into consideration in order to guarantee that the delivered QoS fulfills the terms of the SLA established between the Service Provider and the End-User. As expected, the QoS offered from the Cloud Operator to the Service Provider should be dynamically adapted according to the QoS required by the End-User, the compliance of the Service Provider to the SLA and finally, the QoS perceived by the End-User. The QoS perceived and experienced by the End-User, ultimately composing the QoE, will affect his final satisfaction with the service, and most importantly, form his decision to repurchase the service. This is why the QoE layer must be analyzed in order to control the End-User's satisfaction, and to avoid the churn. To this end, we define a feedback-based convergence process in order to link the different aspects of the QoS/QoE and fulfill all the cloud entities requirements.

[1]Inspired from the pioneering work in [51] and the more recent reference [52].

From the point of view of the Cloud Operator, the main issue resides in how heterogeneous services (e.g., media and ERP) can be hosted together on the same cloud segment (i.e., utilizing the same cloud resources) without being affected by scalability issues (i.e., the number of simultaneous users).

From the Service/Platform Provider's side, there exist two major issues. The first one is to understand the End-Users requirements, to adapt the quality of the service offered according to these requirements, and to establish a feasible SLA. This may become a critical step in cloud computing since negotiating a cloud SLA may involve not only the system/ hardware QoS, supported by the Cloud Operator, but also other contextual aspects like the End-User's expectations or the influence of Service/Platform Provider reputation in clients. The second major issue for the Service/Platform provider deals with the necessary actions to be followed based on the feedback from the End-User. Without a doubt, the actions to be taken should not neglect the QoBiz dimension. For example, the pricing policy should be considered and analyzed, since it may convince the client to migrate its services to the cloud.

Finally, the End-User is concerned about the lack of QoE-related SLAs with the Service/Platform Provider, when considering migrating sensitive services to the cloud. The Service/ Platform Provider is asked to guarantee the operability and the level of quality of the service in a similar way, like being hosted in-house. It is necessary to establish an adaptive mechanism to analyze the QoE and adjust it to enhance the user's satisfaction with the service.

The relations between: (i) the Service/Platform Provider and the End-User, and (ii) the Service/Platform Provider and the Cloud Operator must be mapped into technical/business agreements for all of these purposes, expressed in terms of End-User QoE.

To this end, the novel framework bridge the distance between the End-User and the Service Provider (including the Cloud Operator support) by means of the inter-relation defined between layers, and the feedback-based convergence process. Fig. 3 summarizes the proposed QoE4CLOUD framework.

V. Conclusion

We have reviewed some recent works relevant to providing quality of service guarantees in cloud computing environments. As few studies into QoE-relevant factors exist for cloud service ecosystems, we introduce QoE4CLOUD as the framework that manages quality with unified criteria in clouds. It would be important in future work to carry out experiments to test the usefulness of the management framework in real conditions.

Acknowledgment

Part of this work was funded by the European Union through the SAVASA Project, under grant agreement no 285621.

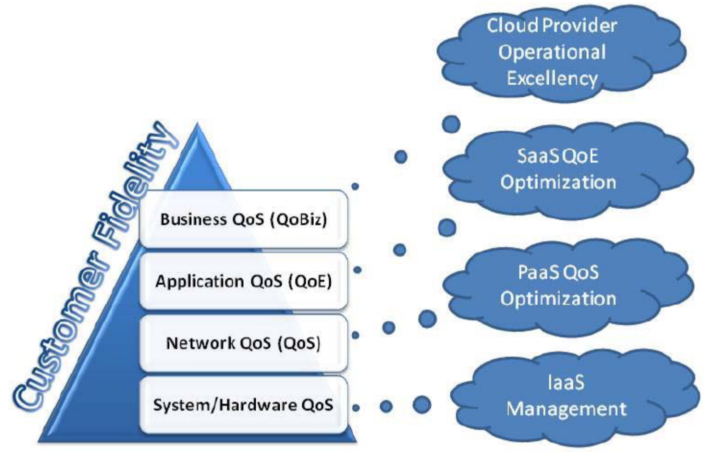

Fig. 3. A layer architecture of the novel QoE4CLOUD approach.

References

[1] "The Future of Cloud Computing. Opportunities for European Cloud Computing Beyond 2010," Expert Group, 2010.

[2] Gartner. [Online]. Available: http://www.gartner.com

[3] IDC. [Online]. Available: http://www.idc.com

[4] NIST Cloud Definition. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[5] CRM - The Entrprise Cloud Computing Company. [Online]. Available: http://www.salesforce.com

[6] Project management software, online collaboration: Basecamp. [Online]. Available: http://basecamphq.com

[7] Google Apps for Business. [Online]. Available: http://www.google.com/apps/intl/en/business/index.html

[8] Windows Azure Platform. [Online]. Available: http://www.microsoft.com/windowsazure

[9] Developer Force: Salesforce.com & Force.com Developer Resources. [Online]. Available: http://developer.force.com

[10] Google App Engine. [Online]. Available: http://code.google.com/intl/el-GR/appengine

[11] Amazon Web Services. [Online]. Available: http://aws.amazon.com

[12] Cloud Computing, Managed Hosting, Dedicated Server Hosting by Rackspace. [Online]. Available: http://www.rackspace.com

[13] Cloud Hosting | Webair Cloud Server Solutions. [Online]. Available: http://www.webair.com/webhosting-cloud-servers.html

[14] V. C. Emeakaroha, M. A. Netto, R. N. Calheiros, I. Brandic, R. Buyya, and C. A. D. Rose, "Towards autonomic detection of SLA violations in Cloud infrastructures," *Future Generation Computer Systems*, vol. 28, no. 7, pp. 1017 – 1029, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X11002184

[15] Cloud monitoring & management. [Online]. Available: http://cloudkick.com

[16] ITSM- network monitoring + service desk management. [Online]. Available: http://www.nimsoft.com

[17] IT Monitoring. [Online]. Available: http://portal.monitis.com

[18] Application and Network Performance with OPNET. [Online]. Available: http://www.opnet.com

[19] Uncover and Resolve Performance Issues with Cloud Performance Monitoring. [Online]. Available: http://www.copperegg.com/product/cloud-monitoring

[20] P. Endo, A. de Almeida Palhares, N. Pereira, G. Goncalves, D. Sadok, J. Kelner, B. Melander, and J. Mangs, "Resource allocation for distributed cloud: concepts and research challenges," *IEEE Network*, vol. 25, no. 4, pp. 42 –46, July-August 2011.

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

[22] S. L. Garfinkel, "An evaluation of Amazons grid computing services: EC2, S3 and SQS," Center for Research on Computation and Society School for Engineering and Applied Sciences, Harvard University, Tech. Rep., 2007.

[23] A. Li, X. Yang, S. Kandula, and M. Zhang, "CloudCmp: shopping for a cloud made easy," in *HotCloud'10*, 2010.

[24] K. V. Vishwanath and N. Nagappan, "Characterizing cloud computing hardware reliability," in *Proceedings of the 1st ACM symposium on Cloud computing*, 2010, pp. 193–204.

[25] P. Bodik, M. P. Armbrust, K. Canini, A. Fox, M. Jordan, and D. A. Patterson, "A Case For Adaptive Datacenters To Conserve Energy and Improve Reliability," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2008-127, Sep 2008. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-127.html

[26] M. Kalyanakrishnan, R. K. Iyer, and J. Patel, "Reliability of Internet Hosts - A Case Study from the End User's Perspective," in *Proceedings of the 6th International Conference on Computer Communications and Networks*, 1997.

[27] F. Panzieri, M. Pellegrini, and E. Turrini, "QoS-aware Clouds," in *Proc. CloudFutures2010*, 2010.

[28] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hossfeld, "An Evaluation of QoE in Cloud Gaming Based on Subjective Tests," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, July 2 2011, pp. 330 –335.

[29] Infovista, "Assuring IaaS Adoption with Customer and Operations Visibility". [Online]. Available: http://www.infovista.com/solutions/iaas

[30] Compuware, "Cloud Performance". [Online]. Available: http://www.compuware.com/application-performance-management/cloud-computing-solutions.htm

[31] J. Hu, J. Gu, G. Sun, and T. Zhao, "A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment," in *Proceedings PAAP '10*, 2010, pp. 89–96.

[32] G. Lodi, F. Panzieri, D. Rossi, and E. Turrini, "SLA-Driven Clustering of QoS-Aware Application Servers," *IEEE Transactions on Software Engineering*, vol. 33, no. 3, pp. 186 –197, March 2007.

[33] V. Stantchev and C. Schröpfer, "Negotiating and Enforcing QoS and SLAs in Grid and Cloud Computing," in *Proceedings of the 4th International Conference on Advances in Grid and Pervasive Computing*, ser. GPC '09, 2009, pp. 25–35.

[34] X. Wang, Z. Du, X. Liu, H. Xie, and X. Jia, "An adaptive QoS management framework for VoD cloud service centers," in *International Conference on Computer Application and System Modeling (ICCASM)*, vol. 1, Oct. 2010, pp. V1–527 –V1–532.

[35] Y. Xiao, C. Lin, Y. Jiang, X. Chu, and X. Shen, "Reputation-Based QoS Provisioning in Cloud Computing via Dirichlet Multinomial Model," in *IEEE International Conference on Communications (ICC)*, May 2010.

[36] Q. Li, Q. Hao, L. Xiao, and Z. Li, "Adaptive Management of Virtualized Resources in Cloud Computing Using Feedback Control," in *Proceedings of the 2009 First IEEE International Conference on Information Science and Engineering*, ser. ICISE '09, 2009, pp. 99–102.

[37] Y. Ye, N. Jain, L. Xia, S. Joshi, I.-L. Yen, F. B. Bastani, K. L. Cureton, and M. K. Bowler, "A Framework for QoS and Power Management in a Service Cloud Environment with Mobile Devices," in *SOSE'10*, 2010, pp. 236–243.

[38] E. Caron, F. Desprez, and A. Muresan, "Forecasting for Grid and Cloud Computing On-Demand Resources Based on Pattern Matching." in *CloudCom 2010*, 2010.

[39] VMware vSphere: Private Cloud Computing for Mid-Size & Enterprises Businesses. [Online]. Available: http://www.vmware.com/products/vsphere/mid-size-and-enterprise-business/overview.html

[40] Microsoft Hyper-V. [Online]. Available: http://www.microsoft.com/servers/hyper-v-server/default.mspx

[41] Red Hat KVM. [Online]. Available: http://www.redhat.com/virtualization/rhev

[42] VMware vCloud Director: Secure Private Clouds, Infrastructure as a Service. [Online]. Available: http://www.redhat.com/virtualization/rhev

[43] Virtualization, Cloud Services, Software - Cloud Computing Platform | Abiquo. [Online]. Available: http://www.abiquo.com

[44] DynamicOps | Operations Virtualization, Cloud Automation, Cloud Management, VDI, Private Cloud. [Online]. Available: http://www.dynamicops.com

[45] Automation, Orchestration, and Provisioning | Gale Technologies. [Online]. Available: http://www.galetechnologies.com

[46] Clusters, Grids, Clouds - Platform Computing. [Online]. Available: http://www.platform.com

[47] C. Weinhardt, A. Anandasivam, B. Blau, and J. Ster, "Business Models in the Service World," *IT Professional*, vol. 11, no. 2, pp. 28–33, 2009.

[48] M. Ajay, G. Zihui, S. Aman, W. Jia, Y. Jennifer, Z. Yin, and Z. Qi, "Towards automated performance diagnosis in a large IPTV-network," in *Proc. ACM SIGCOMM*, 2009.

[49] B. Ali, P. Colin, and O. Jorg, "On the use of RTP for monitoring and fault isolation in IPTV," *IEEE Network Magazine*, vol. 24, no. 2, 2010.

[50] "Performance in the Cloud Survey Report", White Paper Compuware. [Online]. Available: http://www.compuware.com/application-performance-management/cloud-computing-solutions.html

[51] A. V. Moorsel, "Metrics for the Internet Age: Quality of Experience and Quality of Business," in *5th Performability Workshop*, 2001.

[52] R. Stankiewicz, P. Cholda, and A. Jajszczyk, "QoX: What is it really?" *IEEE Communications Magazine*, vol. 49, no. 4, 2011.