International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013
ISSN 2229-5518

Monitoring, Tracking and Quantification of Quality of Service in Cloud Computing

Mohamed Firdhous, Suhaidi Hassan, Osman Ghazali

Abstract— Cloud computing has become a buzzword in computing circles now a days. Due to the attraction of cloud computing, it has attracted several service providers to the market in a very time. These cloud service providers have created a very competitive market for the customers to choose from. Due to the heavy competition, the cost of cloud services must be kept at a minimum in order to attract the sufficient number of customers. Also on the other hand, the service providers must assign as many customers as possible to a single physical system, so that the investment on these systems becomes profitable. When many customers are assigned to a single physical system, the Quality of services (QoS) of the cloud offerings would suffer. Hence it becomes necessary to monitor, track and quantify the QoS of the cloud systems in order to provide the right information to both customers and service providers. This information would help both customers and service providers in terms creating a match between them based on expectations and the capacity to meet them. This would increase the efficiency of the cloud systems by loading them to the optimum levels without sacrificing on the expected quality. Continuous monitoring would help to understand the behavior of the system in the short term and long term helping the service providers to take the necessary remedial actions soon. In this paper, the authors describe the research motivation, objectives, research questions, methodology adopted and significance of the PhD project carried out for developing a QoS monitoring, tracking and quantifying system. It also outlines the progress of the work so far along with the achievements.

Index Terms— Cloud Computing, Quality of Service, Service Level Agreement, Resource Optimization.

---

1 Introduction

Cloud computing has the changed the entire computing landscape by making the resources available over the internet as services. Similar to electricity, water, gas and telephony, computing also becomes a utility under cloud computing [1]. Under the utility computing paradigm, computing resources including hardware, development environment and user applications can be accessed remotely over the Internet and paid for only the usage. In the recent times, due to the popularity of cloud systems the market has been flooded with a large number of cloud service providers [2]. These cloud providers host their services on the Internet and make them available to any customer who would like to purchase them. In [3], Garg, Gopalaiyengar and Buyya state that at any given time, large virtualized systems may host and serve thousands of customers. Though cloud computing systems are advantageous to both customers and service providers in terms of economy and utilization of resources, if the resource provisioning is not carried out optimally it would also become a disaster [4]. Similar to any other subscription based services, prior to the commencement of the services, the service providers and customers enter into an agreement called Service Level Agreement (SLA) [5]. The SLA would contain the roles and responsibilities of the parties involved, scope of services, quality and performance requirements, charges and rates etc. Thus Quality of Services (QoS) plays an important role in making the cloud services acceptable to customers.

In this paper, the authors present a proposal of a project that has been targeted towards designing a secure reliable monitoring, tracking and quantifying system for cloud computing. The paper discusses in detail all the elements of the proposed project namely research motivation, objectives, research problem and questions, methodology adopted, and significance of the work along with the progress so far. The paper also presents a brief literature review that has been carried out as part of this project.

The rest of this paper is organized as follows. Section 2 presents the research motivation that provided impetus for further comprehensive investigation into this exciting field of cloud computing. Section 3 discusses the problem statement and research questions providing a background to the exact research area along with the problems studied. Section 4 details the proposed methodology adopted in this work. Discussion on expected contributions to be made by this PhD research is given in Section 5, while Section 6 provides a brief summary of the related work carried out in this area while Section 7 presents the preliminary research work carried out so far. Conclusion and future work is presented at last in Section 8.

2 Research Motivation

Though cloud computing provides many advantages to both customers and service providers in terms of cost savings and

- *Mohamed Firdhous is a Senior Lecturer attached to the Faculty of Information Technology, University of Moratuwa, Sri Lanka. He is currently on study leave and pursues PhD in computer networks attached to the InterNetWorks Research Lab, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia, E-mail: mfirdhous@internetworks.my, Mohamed.Firdhous@uom.lk*
- *Associate Prof. Dr. Suhaidi Hassan is an associate professor in the School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia. He presently heads the InterNetWorks Research Group and the chairman InterNetWorks Research Lab. Prior to this, he was the Assistant Vice Chancellor of the College of Arts and Sciences, Universiti Utara Malaysia, Malaysia, E-mail: suhaidi@uum.edu.my*
- *Associate Prof. Dr. Osman Ghazali is an associate professor attached to the School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia and a member of the InterNetWorks Research Group. Previously he was the head of the Department of Computer Science, within the School of Computing and Technical Director of the Universiti Teaching Learning Center, Universiti Utara Malaysia, E-mail: osman@uum.edu.my*

utilizations, it still needs to earn the confidence of the customers in certain other aspects for it to become commonly deployed successful technology. Due to the dynamic nature of cloud computing resulting from the creation and hosting of virtual systems on the fly, the performance of the system becomes unpredictable [6]. Many of the commercial applications including multi-tiered business applications, scientific data processing, multi-media applications that can benefit from cloud computing are highly sensitive to quality variations [7]. The QoS requirements to be met by the service provider along with the penalties to be imposed, in case of violation are specified in SLAs signed by both parties [5]. An example SLA to be signed between Amazon Web Services (AWS) a leading cloud provider and its customers who wish to use its Amazon Simple Storage Service (S3) can be found at [8]. As per the SLA, the AWS commits to take commercially reasonable efforts to maintain the availability of Amazon S3 at least at 99.9% during any monthly billing cycle. The compensation for failing to meet the above commitment is service credit also described in the SLA. There is no further commitment made on any other QoS expectations.

From the above discussion, it can be seen that the commitments made by leading cloud service providers at present are too simple and does not mention the complex application specific requirements. This kind of SLAs may not be strong enough to attract business customers whose applications are more sensitive to fluctuations in QoS in many dimensions. The situation has been more aggravated by news item reported in the media about the high-profile crash of Amazon EC2 cloud services [9]. This service outage affected many high profile businesses who had hosted their services at AWS. Not only the site was down for many days and but also some organizations lost their data permanently.

Hence the cloud service providers need to come up with innovative methods to provide the service quality demanded by different types of applications and also to assure them that these commitments will be maintained. Also there should be independent verification of the maintenance of the claims of meeting commitments by service providers. Only if the above can be provided, customers will have confidence on the service providers and would readily move their applications to cloud systems in order to reap the benefits of cloud computing.

The main objective of our research is to come up with an innovative model and mechanisms to monitor track and quantify the dynamically changing the QoS performance of cloud services. The proposed model would be able to track the performance in many dimensions using multiple QoS parameters and quantify them in an easily understandable form. During the request, allocation of resources and executing the required tasks, the performance of the system may face unpredictable challenges due to availability of resources, load, and throughput of hardware services. Hence it is a must to continuously monitor and track the real time QoS performance of cloud systems.

Detecting exceptions, malfunctions and degradations of service quality would help service providers to act proactively and correct them before the systems break down. This would help the service providers to maintain their service quality and confidence cultivated in the customers' minds. System degradations can be detected and handled through the development of an efficient, scalable, interoperable, easy-to-use monitoring tool. In this project our object is to conduct an in depth research in order to achieve the research goals given below.

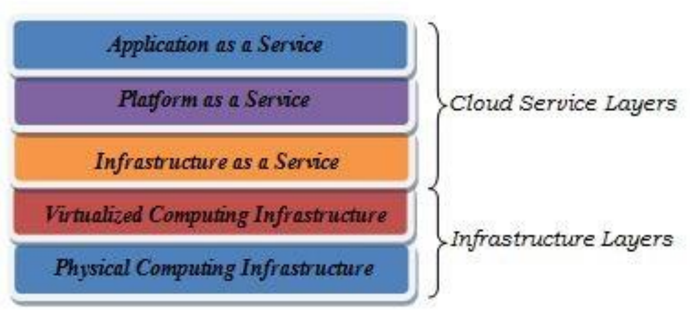
1. To develop an analytical (mathematical) model that can be used to predict the QoS of cloud systems under various conditions.
2. To develop techniques that can dynamically monitor, track and quantify the QoS of cloud computing systems.
3. To develop mechanisms that can distribute the computed QoS score securely among cooperating systems.

3 Problem Statement and Research Questions

The main objective of this research project is to design, develop, implement and test a system that can be used to continually monitor, track and quantify the performance of cloud computing systems. The performance of a cloud computing system is very dynamic due to the very nature of the system itself. Cloud systems create virtual computers and host applications on them on the fly. Similarly they can remove these virtual computers and release the resources back, once the required work has been completed.

Cloud computing services have been divided into three main layers. They are namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [10]. Fig. 1 shows the Cloud services layered model along with the underlying physical computing infrastructure and virtualized computing infrastructure as two distinct layers. The physical hardware is the real workhorse that carries out the processing. The physical hardware is generally provided in the form of computing clusters, grids or individual servers [11]. The virtualized computing infrastructure is created by installing a Virtual Machine Manager (VMM) on the physical hardware [12]. The VMM provides the necessary isolation and security between the multiple virtual machines running in parallel on a single physical computer.

Fig. 1: Cloud Computing Layered Model



IaaS is the provision of virtual hardware as a service over the Internet. These virtual machines can be brought up and removed on the fly based on customer demand. Once a virtual machine has been purchased, it can be treated as if it is real hardware and any operating system and applications can be installed on it. PaaS is the complete software development environment along with operating system, development and testing tools and application programming interface installed on virtual hardware. PaaS helps web based application developers to reduce the cost and time of bringing their applications to market from the design boards. SaaS is the new paradigm of software marketing and ownership. SaaS enables customers to access web based applications hosted in remote data centers and pay only for the usage. These applications have the capability of managing their own data and configuration information to suit individual user requirements.

From the above description of cloud computing systems, it can be seen that cloud systems can host applications and services that have drastically different requirements in terms of QoS [13]. While the transactional applications demand better response times and throughput guarantees, the non-interactive batch jobs more are concerned with job completion times and accuracy of processing [14]. Thus, it can be concluded that the QoS demands of cloud services are more complex and depends on multiple factors or parameters.

This research would be designed in such a manner to find answers to the following research questions.

RQ1: How to design an analytical (mathematical) model that can be used to predict the QoS behavior of the cloud system under different conditions?
RQ2: How to develop techniques that can dynamically monitor, track and quantify the QoS of cloud computing systems?
RQ3: How to devise mechanisms that can distribute the computed QoS scores among cooperating systems?

RQ1: How to Design an Analytical Model?

In a shared distributed dynamic system like cloud computing, it is important that the systems must be loaded appropriately to achieve the competing requirements higher revenue for service providers and better services for customers. In order to achieve the higher revenue, the systems must be loaded optimally without sacrificing the quality of service performance. If the service providers have a prior knowledge of optimum load levels and the cost of the resources, they can price the services optimally that ensures both profits and maintains quality of services. The challenge in developing analytical models for predicting the behavior of complex dynamic systems is fitting the right statistical models to the different components of the systems such as request arrival pattern, service time distributions, I/O system behaviors, failure handling, resource usage etc. The challenge is further complicated due to interdependence of components with each other. Hence, there is an urgent need for a cloud service performance modeling and workload prediction technique that can ensure optimum system utilization without sacrificing the QoS requirements. RQ1 that has been formulated in order to solve handle this practical issue is: Is it possible to develop a mathematical model that can effectively predict the behavior of cloud system under various conditions specified by different QoS metric values?

RQ2: How to Develop Monitoring Techniques?

A cloud computing system is a very dynamic one compared to other distributed systems such as cluster computing, grid computing etc. The dynamic creation and hosting of virtual systems aggravates the situation due to non commitment of any resources until such a virtual system becomes active. Also cloud system can host services at different layers of abstraction and applications that have distinct QoS demands based on different metrics. Hence it is necessary to continuously monitor the performance and track the changes in them. Also, it is important to protect the system from malicious attacks and momentary fluctuations. The research question, RQ2 formulated for meeting this challenge is: Is it feasible to design a secure reliable technique that can monitor, track and quantify the performance of cloud systems?

RQ3: How to Devise Score Distribution Mechanisms?

It is a common practice for service providers to host their systems to geographically distributed data centers. Distributing data centers geographically across multiple sites help the service providers to handle short term and long term disasters gracefully. Though the systems are distributed across multiple sites, it must be transparent to the end users and behave like a single system. When monitoring such a distributed system, it is not practically advisable to have a single monitoring system for all the sites. When multiple systems are employed, they should cooperate with each other sharing their information. When scores are shared among geographically distributed systems, they should consider other factors such as network dynamics when updating the scores and security of transmitted/received information. RQ3, the research question that was developed for researching into this challenging area is: Is it practical to devise a secure reliable mechanism that can cooperatively exchange QoS scores among themselves and incorporate the other information such as network dynamics into those scores?

4 METHODOLOGY

The main objective of this research is to develop a reliable and secure monitoring system for cloud computing performance. In order to achieve this objective, it is necessary to follow a strict scientific procedure as the results obtained as valid and repeatable under similar conditions. The methodology adopted for carrying out this research consists of five main phases. They are namely, analysis, design, testing, verification & validation and implementation. Fig. 2 shows these phases and how there are connected to each other in a graphical format.

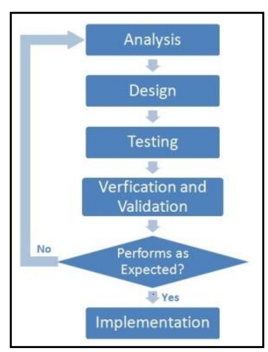


Fig 2: Research Methodology

The project will be divided into three main tasks in finding solutions to the three research questions identified in Section 3. The tasks are briefly described below.

Task 1: Designing an Analytical Model for Cloud Computing

Modeling of cloud computing mathematically enables researchers and other professionals to carry out in depth analyses of the system under different conditions. In solving RQ1, it expected to design and develop an analytical (mathematical) model that would correctly predict the behavior of the cloud computing system under various conditions. The model is expected to be based on computational statistical techniques as the behavior of components in a cloud system can be treated as stochastic processes for all practical purposes [15]. These models will capture the behavior of different components in terms of QoS attributes such as time, throughput, utilization, cost etc.

Task 2: Developing Techniques for Monitoring Cloud Performance

Task 2 would mainly concentrate on how to monitor the behavior of a cloud system in terms of meeting the QoS requirements of customers and converting it to a comprehensible score. The proposed techniques must be capable of monitoring the QoS performance of the cloud system based on more than one parameter and be able to assign relative preference to these metrics depending on the user requirements. The proposed system should continuously monitor the performance and update the final score as these systems are dynamic. Also the proposed techniques must be able to identify and isolate the deviations in performance based on whether it was due to momentary fluctuations in performance or due to permanent degradations. For the proposed techniques to be successfully employed in practical systems, it should have enough resilience to attack that aims to modify the scores maliciously. Hence the ultimate mechanism that would be developed in solving RQ2 would consist of multiple components where each one would work independently towards achieving the final goal of monitoring, tracking and quantifying the service quality of cloud system effectively and securely.

Task 3: Devising Mechanisms for Distributing Performance Scores

Task 3 is mainly concerned with the collaboration between independent cloud monitoring systems developed in Task 2. In a cloud system that has been deployed across a wide geographical area and also services globally dispersed clients must be identify and allocate the right resources hosted at the right locations. Also it is practically not feasible to monitor and track all the systems with a single monitoring system. Hence it necessary to deploy multiple systems at various locations, so that they can monitor and track the performance of the systems independently but collaboratively exchange the information collected with each other. When the information is exchanged, they should also consider the other intervening factors that can affect the service quality such as the quality and performance of the network connecting these systems together. The other major factor that needs to be taken into account is the security of the information transmitted. The transmitted information can be attacked enroute by malicious attackers or the systems receiving the information may be fed with wrong information. Hence developing mechanism for exchange of information between collaborating monitoring units in answering RQ3, it is necessary arrive at an optimum solution that considers all the factors in order to arrive at a resilient, secure and scalable mechanism.

5 Significance

The significance of the proposed project is manifold. The project is expected to have three main individual but related contributions. These contributions would help greatly for enhancing acceptability of cloud computing to a wider audience. Contribution 1, the analytical model would help the network designers to prepare their resources in such a manner that is most suitable to meet the requirements of the customers. This model can also be used by researchers to analyze the behavior of cloud systems under different conditions. Contributions 2 and 3 together would help the customers and service providers equally. From the customers' point of view, the monitoring, tracking and quantifying unit would help them to identify the service providers who would likely to meet their QoS requirements. From the service providers' angle, this system would help them to track the performance of their resources. The service providers would be able to identify any degradation of performance well before it becomes a disaster and irreversible.

6 Related Work

This section briefly discusses the related work that has been carried out by other researchers and reported as published work in conferences and journals. When selecting the literature for reviewing, special attention was paid for selecting the papers that were relevant and most recent. Instead of just listing the work, a critical analysis on these proposed mechanisms was carried out with special reference to the principles, strengths and weaknesses.

Cloud computing systems may host thousands of globally dispersed clients at any given time. These clients may access different types of services that have varying requirements depending on the type of clients, services and resources involved. In order to meet the requirements of clients and services, it is necessary to provide a certain level of QoS by the service providers. Nevertheless, providing a guaranteed QoS in such a challenging environment in a widely distributed diverse networks supporting complex hosting of services is not an easy task [16,17]. Though it is a challenging task, several researchers have undertaken to develop mechanisms, frameworks and systems which could guarantee the QoS requirements of different services. This section takes an in depth look at these mechanisms, frameworks and systems.

In [18], Liu et al have proposed a generic QoS framework for cloud workflow system. The proposed framework covers all the four stages of cloud workflow namely, QoS requirement specification, QoS-aware service selection, QoS consistency monitoring and QoS violation handling. The shortcoming of this framework is that it does not specifically identify any QoS parameters and also does not discuss how to differentiate clients requiring different QoS levels.

Chen and Zhang have proposed a workflow scheduling algorithm based on Particle Swarm Optimization (PSO) in [19]. The proposed mechanism can optimize up to seven parameters specified the users compared to traditional optimization techniques that consider only the workflow execution time. The downside of the proposed mechanism is that it lacks a monitoring scheme for catching QoS violations or punishing the violators.

Buyya, Garg, and Calheiros have proposed a framework for SLA management with special reference to managing QoS requirements in [17]. The proposed architecture successfully integrates the market based resource provisioning with virtualization technologies for flexible resource allocations to user applications. But the proposed architecture does not support different cloud service offerings such as IaaS, PaaS and SaaS together in a combined manner.

In [5], Feng et al have proposed an optimal resource allocation model for revenue maximization. The model has been mathematically derived and tested using both synthetic and traced datasets. The proposed model performs better than heuristic optimization of resources in maximizing profits. But the application of this method is limited as it considers only the mean response time as the QoS attribute to be satisfied. For customers who require guaranteed performance or at least a commitment in terms of a confidence level cannot be served through this model. Hence from the customers' point of view, the model has limited application and may serve only casual users.

In [20], den Bossche, Vanmechelen and Broeckhove have proposed a set of heuristics for scheduling deadline-constrained applications in a hybrid cloud system in a cost effective manner. The proposed system attempts to maximize the use of local resources along with minimizing the use of external resources without compromising the QoS requirements of the applications. The optimization heuristics takes the cost of both computation and data transfer along with the estimated data transfer times. The main criteria in optimization is the maximization of cost saving. The effect of different cost factors and workload characteristics on the cost savings have been analyzed along with the sensitivity of the results to the different runtime estimates. The advantages of the proposed methodology is that it can select an optimized set of resources from both in-house (private) and public cloud systems for meeting the QoS requirements. But at the same time it suffers from certain weaknesses. Though it is concerned only about the deadline concerned applications, it does not consider the failures that may occur after the scheduling has been done. The failure will increase the cost of execution and affect the application in terms of quality.

In [21], Emeakaroha et al have presented a scheduling heuristic that takes multiple SLA parameters when deploying applications in the Cloud. The attributes considered are physical requirements such as CPU time, network bandwidth and storage capacity for deploying applications. These parameters have limited application in real world systems as they need to be considered only during deployment. Once the applications have been ready for client access, the customers would be more interested in performance parameters such as response time, processing time etc. Hence this heuristic may not have much practical significance in real world business environments.

Li et al in [22] have proposed a novel customizable cloud workflow scheduling model. The authors have incorporated trust into the model in addition to the QoS targets. In order to analyze the users' requirements and design a customized schedule, the authors propose a two stage workflow model where the macro multi-workflow stage is based on trust and micro single workflow stage classifies workflows into time-sensitive and cost-sensitive based on QoS demands. The classification of workflows has been carried out using fuzzy clustering technique. The proposed model restricts the QoS parameters considered to response time, bandwidth, storage, reliability and cost. Also the delivery of QoS is confined only to average values and no guarantee of service delivery is provided at least in terms of a predetermined confidence level. This is a strong limitation of the proposed technique as the users do not have the freedom to select their own QoS parameters and no guarantee of the QoS delivery at the least a statistical validation.

Alhamazani et al., in [23] have outlined the importance of dynamically monitoring the QoS of virtualized services. they further claim that the monitoring of the services would help both the cloud provider and application developer to maximize the return of their investments in terms of keeping the cloud services and hosted applications operating at peak efficiency, detecting changes in service and application performance, SLA violations, failures of cloud services and other dynamic configuration changes. The paper mainly concentrates on describing the PhD work being carried out in terms of research questions, objectives and methodology. The researchers mainly concentrate on SNMP based QoS monitoring. Since this is a concept paper describing work in progress, no concrete proposal is put forward or evaluated.

The literatures discussed above are mainly concerned with cloud workflow. The cloud workflows attempt to select the resources in such a manner that the required QoS would be satisfied. None of the literature cited above discuss continuous monitoring of cloud systems for their performance or quantifying them. Also, the reported mechanisms are unable to identify or detect system degradations as they are mainly concerned with resource selection and allocation.

7 Progress Till To-date

The project has progressed significantly and the tasks identified in Section 4 have been carried out in parallel. The development of the cloud computing model has been almost complete and requires slight modifications in terms of performance tuning and testing. The development of monitoring and tracking system proposed to be carried out under Task 2 has progressed significantly. This work has been carried out iteratively by improving a basic model designed at the beginning until the final goal of robust mechanism has been achieved. The basic design and improvements have been presented at various international conferences and forums. These designs received very positive and encouraging comments from reviewers who reviewed these works.

8 Conclusions and Suggestions for Future Work

Cloud computing systems have become very popular in the recent times and attracted the attention of many people including researchers, service providers and customers. The cloud systems provide many advantages over the traditional computing system due to the innovative way of making the computing resources available over the internet and charging the customers. Cloud systems employ a pay-as-you-go business model similar to utilities like electricity, water, gas and telephony. Quality of Service would play an important role in making cloud computing acceptable to everyone especially the business customers.

Monitoring the cloud system is a key factor in ensuring the committed service quality is maintained. The monitoring the system will help both the customers and service providers as the customers can select the right service provider who could meet their requirements and service providers would be able design manage their systems optimally meeting the requirements of the customers.

The conclusion of the proposed study, it is expected to contribute significantly to the existing knowledge on cloud computing with special reference to enhancing the service quality. The work is also significant practically as the systems once completed can be used by both customers and service providers to obtain a better service and enhance their services and profitability respectively.

References

[1] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," *Journal of Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, June 2009.

[2] B.P. Rimal, E. Choi, and I. Lumb, I. "A Taxonomy and Survey of Cloud Computing Systems," *Proc. 5th Int. Joint Conf. INC, IMS and IDC (NCM'09)*, pp. 44-51, 2009.

[3] S.K. Garg, S.K. Gopalaiyengar, and R. Buyya, "SLA-based Resource Provisioning for Heterogeneous Workloads in a Virtualized Cloud Datacenter," *Proc. 11th Int. Conf. Algorithms and Architectures for Parallel Processing (ICA3PP2011)*, pp. 371–384, 2011.

[4] D. Bourcier, and P.D. Filippi, "Cloud Computing: New Research Perspectives for Computers and Law," *Proc. 13th Int. Conf. Artificial Intelligence & Law*, pp. 73–92, 2011.

[5] G. Feng, S. Garg, R. Buyya, and W. Li, "Revenue Maximization using Adaptive Resource Provisioning in Cloud Computing Environments," *Proc. 13th ACM/IEEE Int. Conf. Grid Computing,* pp. 192–200, 2012.

[6] D.C. Marinescu, *Cloud Computing: Theory and Practice*. Burlington, MA: Morgan Kaufmann, 2013.

[7] A. Iosup, N. Yigitbasi, and D. Epema, "On the Performance Variability of Production Cloud Services," *Proc. 11th IEEE/ACM Int. Symp. Cluster, Cloud and Grid Computing,* pp. 104-113, 2011.

[8] Amazon S3 Service Level Agreement, http://aws.amazon.com/s3-sla/. Accessed on February 07, 2013.

[9] H. Blodget, "Amazon's Cloud Crash Disaster Permanently Destroyed Many Customers' Data." http://www.businessinsider.com/amazon-lost-data-2011-4. Accessed on February 07, 2013.

[10] C. Vecchiola, S. Pandey, and R. Buyya, "High-performance cloud computing: A view of scientific applications," *Proc. 10th Int. Symp. Pervasive Systems, Algorithms, and Networks (ISPAN),* pp. 4–16, 2009.

[11] R. Buyya, A. Beloglazov, and J.H. Abawajy, "Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges", *Proc. Int. Conf. Parallel and Distributed Processing Techniques and Applications,* pp. 6-20, 2010.

[12] Y. Li, W. Li, and C. Jiang, "A Survey of Virtual Machine System: Current Technology and Future Trends," *Proc. Third Int. Symp. Electronic Commerce and Security,* pp. 332-336, 2010.

[13] A. Quiroz, H. Kim, M. Parashar, N. Gnanasambandam, and N. Sharma, "Towards Autonomic Workload Provisioning for Enterprise Grids and Clouds," *Proc. 10th IEEE/ACM Int. Conf. Grid Comp.,* pp. 50-57, 2009.

[14] D. Carrera, M. Steinder, I. Whalley, J. Torres, and E. Ayguade, "Enabling resource sharing between transactional and batch workloads using dynamic application placement," *Proc. 9th ACM/IFIP/USENIX Int. Conf. Middleware,* pp. 203-222, 2008.

[15] T.M. Chen, "Network Traffic Modeling," *The Handbook of Computer Networks: Distributed Networks, Network Planning, Control, Management, and New Trends and Applications, Vol. 3,* H. Bidgoli, ed., Hoboken, NJ: Wiley, pp. 326-339, 2007.

[16] J.M. Pedersen, M.T. Riaz, J.C. Júnior, B. Dubalski, D. Ledzinski, and A. Patel, "Assessing Measurements of QoS for global Cloud Computing Services," *Proc. Ninth IEEE Int. Conf. Dependable, Autonomic and Secure Computing,* pp. 682-689, 2011.

[17] R. Buyya, S.K. Garg, and R.N. Calheiros. "SLA-Oriented Resource Provisioning for Cloud Computing: Challenges, Architecture, and Solutions," *Proc. Int. Conf. Cloud and Service Computing,* pp. 1-10, 2011.

[18] X. Liu, Y. Yang, D. Yuan, G. Zhang, W. Li, and D. Cao, "A Generic QoS Framework for Cloud Workflow Systems," *Proc. Ninth IEEE Int. Conf. Dependable, Autonomic and Secure Computing,* pp. 713-720, 2011.

[19] W.N. Chen, and J. Zhang, "A Set-Based Discrete PSO for Cloud Workflow Scheduling with User-Defined QoS Constraints," *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, pp. 773-778, 2012.

[20] R.V. den Bossche, K. Vanmechelen, and J. Broeckhove, "Cost-Efficient Scheduling Heuristics for Deadline Constrained Workloads on Hybrid Clouds," *Proc. 3rd IEEE Int. Conf. Cloud Comp. Tech. and Sc., (CloudCom),* pp. 320-327, 2011.

[21] V.C. Emeakaroha, I. Brandic, M. Maurer, and I. Breskovic, "SLA-Aware Application Deployment and Resource Allocation in Clouds," *Proc. 35th IEEE Annual Computer Software and Applications Conference Workshops (COMPSACW),* pp. 298-303, 2011.

[22] W. Li, Q. Zhang, J. Wu, J. Li, and H. Zhao, "Trust-based and QoS Demand Clustering Analysis Customizable Cloud Workflow Scheduling Strategies," *Proc. IEEE Int. Conf. Cluster Comp. Workshops,* pp. 111—119, 2012.

[23] K. Alhamazani, R. Ranjan, F. Rabhi, L. Wang, and K. Mitra, "Cloud Monitoring for Optimizing the QoS of Hosted Applications," *4th IEEE Int. Conf. Cloud Comp. Tech. and Sc.*, pp. 765—770, 2012.