A SURVEY ON QUALITY OF SERVICE IMPLEMENTATIONS IN CLOUD COMPUTING

Merly Mathew

P G Student, CSE, LBSITW, Kerala (India)

ABSTRACT

Cloud computing is a new terminology achieved by distributed, parallel and grid computing and a design pattern for large, distributed data centers. Cloud computing offers end customers a pay as go model. Quality of service plays an important factor in distributed computing. Cloud computing provides different types of resources like hardware and software as service via internet. Under cloud computing, computing resources are hosted in the internet and delivered to customers as services. Prior to that, the customers and cloud provider negotiate and enter into an agreement named service level agreement. The service level agreements clarify the roles, set charges and expectations and provide mechanisms for resolving service named problems within a specified and agreed upon time period. Service level agreements also cover performance, reliability conditions in terms of quality of service guarantees. In this paper, the authors present a comprehensive survey on quality of service implementations in cloud computing with respect to their implementation details, strengths and weaknesses.

Keywords: Cloud Computing, QoS, Scheduling, SLA, VMM

I. INTRODUCTION

Cloud computing is the 5th utility after electricity, water, gas and telephony. Nowadays, the market has been flooded with a large number of cloud service providers. These eservices are hosted on internet and is available to customer who wants to purchase it. In terms of economy and resource utilization, the cloud computing is advantageous to both customers and service providers but if optimal resource utilization is not carried out, it would become a disaster. Prior to commencement of services both service providers and customers enter into an agreement called Service Level Agreement(SLA), which contain the roles and responsibilities of both parties, scope of services, quality and performance requirements, charges and rates. Thus Quality of Services (QoS) plays an important role in making cloud service acceptable to customers. In this paper, a survey on mechanisms and methods proposed by various researchers with respect to their implementation principles, strengths and weakness is carried out.

II.CLOUD COMPUTING

The main characteristics of Cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. The cloud model is composed of five essential characteristics, three service models, and four deployment models.

2.1. Essential Characteristics

- On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
- Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
- Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.
- Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts).

The service of the cloud computing is divided into three main categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). Fig.1 shows the Cloud computing layers along with the underlying physical computing infrastructure and virtualized computing infrastructure as two distinct layers [24]. 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. The virtualized computing infrastructure is created by installing a Virtual Machine Manager (VMM) on the physical hardware. The VMM provides the necessary isolation and security between the multiple virtual machines running in parallel on a single physical computer.

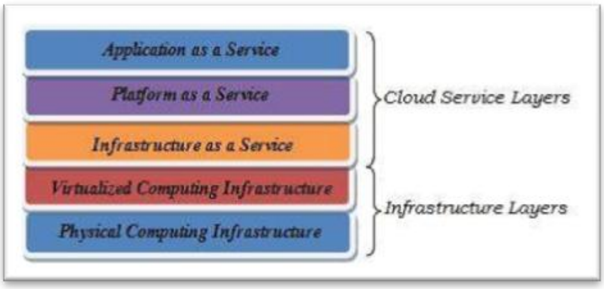
Application as a Service
Platform as a Service
Infrastructure as a Service
Virtualized Computing Infrastructure
Physical Computing Infrastructure
Cloud Service Layers
Infrastructure Layers

Fig.1: Cloud Computing Layers

III.RELATED WORKS

The different QoS parameters considered in various experiments are CPU time, network bandwidth, storage capacity, response time, performance time, processing time. Table 1 summarizes the work done so far with reference to their strengths and weaknesses along with the proposed model or framework. From Table 1, it can be seen that there is still a lot of scope for future work in this area.

Table 1: Summary of Strengths and Weakness of Proposed Models and Frameworks

Work	Proposed Model/framework	Strengths	Weakness
[1]	A framework for SLA management with special reference to managing QoS requirements.	Successfully integrates the market based resource provisioning with virtualization technologies for flexible resource allocations.	Does not integrate IaaS, PaaS and SaaS in a combined manner.
[2]	A generic QoS framework for cloud workflow	Covers all the four stages of cloud workflow.	QoS metrics are not identified and no mechanism for differentiating customers based on requirements.
[3]	A set-based PSO approach scheduling problem in cloud computing.	Multiple parameter optimizations are possible.	But no monitoring mechanism is implemented for catching violations.
[4]	A set of heuristics for scheduling deadline-constrained applications in a hybrid cloud system.	The optimization heuristics takes the cost of both computation and data transfer along with the estimated data transfer times and different cost factors and workload characteristics.	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.
[5]	A scheduling heuristic that takes multiple SLA parameters when deploying applications in the cloud	Considers deployment attributes such as CPU time, network bandwidth, storage capacity etc, before installation of applications in the cloud system.	Does not consider performance parameters such as response time, performance time etc.
[6]	A flexible multistage work-flow scheduling model.	The proposed model is flexible due to breaking up of the workflow scheduling mechanism into multiple stages and grouping the requests based on the user requirements.	Application is strongly limited due to strict restriction on the type of QoS attributed taken into account and the absence of QoS delivery guarantees.
[7]	The correlation between QoS/QoE has been studied.	QoS/QoE correlation has been studied using a selected set of machine learning techniques.	Discuss more about the capabilities of machine learning techniques than about QoS or QoE. The QoS/QoE correlation is a

			case for evaluating the machine learning techniques.
[8]	Proposal for monitoring the cloud system for QoS performance	Only the concept and idea based work in progress have been described.	No concrete proposal or evaluation is presented.
[9]	Profit-Based Analysis of Resource Allocation on QoS	An innovative method for analyzing the impact of resource provisioning.	No discussion on how to optimally allocate resources.
[10]	A distributed resource allocation algorithm for cloud and grid systems.	Capable of handling multiple resource requirements.	Too simple, as it assumes perfect conditions for execution. Failures after allocation of resources are not taken into account.
[11]	Extensible dynamic provisioning framework for multitenant cloud system.	The proposed framework is dynamic and allocates resources depending on the tenant requirements.	May not be capable of handling bursty requirements with short duration and large resource requirements. The new tenants arriving late may suffer from resource starvation.
[12]	Lightweight framework for monitoring public clouds.	Less resource intensive	Does not monitor the real QoS parameters such as response time, processing time etc.
[13]	A framework for handling adaptive applications in cloud systems.	Based on multi-input multi-output feedback control model for resource provisioning	Limited only to CPU and memory provisioning. Hence application performance may be affected by other resource constraints such as network, storage etc.
[14]	A resource pricing model for QoS and profit balancing.	Uses realistic values using age as a parameter.	Utilization is not considered in computing cost. Hence may produce inaccurate costs.
[15]	A monitoring application for	Can be used by clients to monitor	Very narrow application

	QoS parameters in iO55.	the performance of service providers.	due to focusing only on available transfer rate and one-way delay as QoS parameters.
[16]	A QoS based trust management model.	Multiple QoS parameters can be used.	No clear explanation on how to use the parameters is given nor is there any possibility to prioritize the parameters.
[17]	Resource allocation in a Compute Cloud through bargaining approach.	The proposed strategy handles the dynamic nature of cloud very well during run time.	May lead to sub optimal solutions from a customer's perspective, if a single provider cannot meet all the requirements.
[18]	Investigation of the capability of MAP based queuing models for predicting workload of cloud systems.	Markov arrival processes have the capability fir heavy trial distributions that are common in web applications.	Only numerical experiments have been used to validate the model, hence needs further validation with real data traces.
[19]	An optimization framework for cross layer cloud services.	Suitable for vendors selling products across multiple layers. Dynamic nature of cloud has been considered.	Lacks the run time management of QoS performance.
[20]	Algorithms for resource allocation for SaaS providers for balancing cost and QoS.	It helps reduce the cost of SaaS providers without compromising the QoS of customers.	Due to reuse of already open VMs, it can create security problems for customers.
[21]	Results of an initial investigation of using Dwarf bench-marks to measure the performance of virtualizes]d hardware.	General labeling of cloud service providers for size or the number of units used is not sufficient to predict the real capabilities through real experiments.	A set of experiments by experts in laboratory may not help the general set of customers who are not that tech sawy.
[22]	A process for matching providers' capability with customers' requirements based on SLA parameters.	Automates the matching process that was hitherto done manually by customers.	Match capabilities published by service providers with customer requirements. It cannot track the changes in cloud

			performance due to dynamic nature of clouds.
[23]	Optimal resource allocation model for revenue maximization.	Mathematically derived and performs better than heuristics.	Only mean performance time is considered, hence not suitable for QoS sensitive applications requiring guaranteed performance.

IV. CONCLUSION

Cloud computing has been the paradigm shift in distributed computing due to the way the resource is provisioned and charged. Managing QoS is a critical task in making such an innovative technology to a larger audience. Several researchers have put forward their ideas for new and innovative solutions for handling this vital area. In this paper, a critical review of the most recent work carried out in this area is done. The findings in terms of the strengths and weaknesses of the pro-posed work have been presented in a table for easy reference.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] M.S.Mushtaq, B.Augustin, and A.Mellouk, “Empirical Study based on Machine Learning Approach to Assess the QoS/QoE Correlation”, Proc. 17th European Conf. Networks and Optical Comm., pp. 1-7, 2012.

[8] K. Alhamazani, R. Ranjan, F. Rabhi, L. Wang,and K. Mitra, “CloudMonitoring for Optimizing the QoS of Hosted Applications,” Proc. 4th IEEE Int. Conf. Cloud Comp. Tech. and Sc., pp. 765-770, 2012.

[9] J. Li, Q. Wang, D. Jayasinghe, S. Malkowski, P. Xiong, C. Pu, Y. Kanemasa, and M. Kawaba, "Profit-Based Experimental Analysis of IaaS Cloud Performance: Impact of Software Resource Allocation," Proc. Ninth IEEE Int. Conf. Services Computing, pp. 344-351, 2012.

[10] D.Adami, C.Callegari, S.Giordano, and M.Pagano, "A Hybrid Multidimensional Algorithm for Network-aware Resource Scheduling in Clouds and Grids," Proc. IEEE Int. Conf. Comm., pp. 1297-1301, 2012.

[11] A. Gohad, K. Ponnalagu,and N.C.Narendra, "Model Driven Provisioning in Multi-tenant Clouds." Proc. Annual SRRI Global Conference, pp. 11-20, 2012.

[12] J. Ma, R. Sun, and A.Abraham, "Toward a Lightweight Framework for Monitoring Public Clouds," Proc. Fourth Int. Conf. Computational Aspects of Social Networks, pp. 361-365, 2012.

[13] Q. Zhu and G. Agrawal, "Resource Provisioning with Budget Constraints for Adaptive Applications in Cloud Environments," IEEE Trans. Services Computing, vol. 5, no. 4, pp. 497-511.

[14] B. Sharma, R.K. Thulasiram, P. Thulasiraman, S.K. Garg, and R. Buyya, "Pricing Cloud Compute Commodities: A Novel Financial Economic Model," Proc. 12th IEEE/ACM Int. Symp. Cluster, Cloud and Grid Computing, pp. 451-457, 2012.

[15] F. Stoicuta, I. Ivanciu, E. Minzat, A.B. Rus, and V. Dobrota, "An OpenNetInf-Based Cloud Computing Solution for Cross-Layer QoS: Monitoring Part Using iOS Terminals," Proc. 10th Int. Symp. Electronics and Telecomm., pp. 167—170, 2012.

[16] M.K. Goyal, A. Aggarwal, P. Gupta, and P. Kumar, "QoS based Trust Management Model for Cloud IaaS," Proc. 2nd IEEE Int. Conf. Parallel, Distributed and Grid Computing, pp. 843-847, 2012.

[17] G.N. Iyer and B, Veeravalli, "On the Resource Allocation and Pricing Strategies in Compute Clouds using Bargaining Approaches," 17th IEEE Int. Conf. Networks, pp. 147-152, 2011.

[18] S.P. Sanchez, G. Casale, B. Scotney, S. McClean, G. Parr,and S. Daw-son, "Markovian Workload Characterization for {QoS} Prediction in the Cloud," Proc. IEEE Int. Conf. Cloud Computing, pp. 147-154, 2011.

[19] Y. Kouki, T. Ledoux, and R. Sharrock, "Cross-Layer SLA Selection for Cloud Services," Proc. First Int. Symp. Network Cloud Computing and Applications, pp. 143-147, 2011.

[20] L. Wu, S.K. Garg, and R. Buyya, "SLA-based Resource Allocation for Software as a ServiceProvider (SaaS) in Cloud Computing Environments," Proc. 11th IEEE/ACM Int.Symp. Cluster, Cloud and Grid Com-puting,pp. 195-204, 2011.

[21] S.C.Phillips, V.Engen, and J. Papay, "Snow White Clouds and the Seven Dwarfs," Proc. Third IEEE Int. Conf. Cloud Comp. Tech. and Sc.,pp. 738-745, 2011.

[22] T. Chauhan, S. Chaudhary, V. Kumar,and M. Bhise, "Service Level Agreement Parameter Matching in Cloud Computing," World Cong. ICT, pp. 564-570, 2011.

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

[24] Mohamed Firdhous, Suhaidi Hassan, Osman Ghazali, "A Comprehensive Survey on Quality of Service Implementations in Cloud Computing", International Journal of Scientific & Engineering Research, Volume 4, Issue 5., pp. 118-123, 2011.