Review on Various Energy-Aware and SLA Violation Reduction Techniques

Praveen Shukla
M.Tech Scholar
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, 462051, India

R. K. Pateriya
Associate Professor
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, 462051, India

ABSTRACT
Whenever large number of demands arrive at the cloud service provider (CSP) it is unable to provide the services as mentioned in service level agreement (SLA) to the cloud customers. This is due to under provisioning of resources resulting when CSP want to earn more profit with limited amount of resource. Cloud service provider accommodate lots of servers, some of them are either overloaded server or underloaded server. Host overloading may cause increase in VM migration and SLA violation. In Cloud environment, to reduce SLA violation there are numerous techniques available but cloud provider select those which are best suited for application requirement and user demand. This paper presents detailed review on the various SLA violation reduction techniques.

Keywords
Cloud service provider (CSP), SLA violation.

1. INTRODUCTION
In today's era of computing, cloud computing emerges as powerful computing model among the various computing models like grid computing, cluster computing etc. Cloud computing provides online storage of data, which can be accessible from multiple distributed and connected resources. Cloud computing depends upon the sharing of physical and/or virtual resources, rather than deploying new hardware and software.

Cloud computing can be visualized as a new outsource service model where cloud service provider outsource their resources on pay as per-use basis. Hence this model benefits both parties (cloud service provider and clients).Everything is available on demand, no need to own things, any time you can access service of cloud as per your request.

Cloud computing is an abstract model where user don't know about data, how and where it is organized and what platform and hardware configuration isbeing used to deliver the services. Hence the underlying architecture is abstract to the user [1].

2. TYPES OF CLOUD
Cloud can be classified on basis of the owner of cloud data center. There are four different types of cloud which are listed as follows:

2.1 Private cloud
It is cloud infrastructure which is available for a single organization. A Private cloud depends upon the virtualization of an organization framework [2].Private cloud is not shared with other organization. It can be maintained by itself or by a third party and it can be hosted internally or externally.

Internal Private Cloud is hosted within own premises. In this type of cloud organization has full control over data but it incurs a maintenance burden. Whereas externally hosted private cloud are also exclusively used by one organization, but are hosted by a third party specializing in cloud infrastructure. As compared to public clouds private clouds are more robust and costly [3].

2.2 Public cloud
Itis cloud infrastructure which is hosted by a third party vendor where services are available for a multiple organization over internet. Generally, public cloud providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access over the Internet [4].

In public cloud customer don't have any idea about the platform (such as Oracle, MySQL etc) use to run the data, where the infrastructure is located. The sharing of cloud infrastructure among the multiple cloud customers provides low cost services via pay as you go model. Inspite of that public cloud provider high efficiency but it is also more vulnerable than private clouds.

2.3 Hybrid cloud
It is a combined infrastructure of two or more cloud (Private cloud, public cloud and community cloud). It provides a flexibility to run some application in private cloud while others in public cloud.

It is more sophisticated way of providing public cloud to your consumer and private cloud for internal organization. The downside is that you have to keep track of multiple cloud security platforms and ensure that all aspects of your business can communicate with each other [4].

2.4 Community cloud
Itis a collaborative environment in which sharing of computing infrastructure in between organizations of the same community. Community such as government organizations within the state of India may share computing resource on the cloud to manage data related to peoples of India.Maintenanceof cloud infrastructure would be done by itself or by a third party.

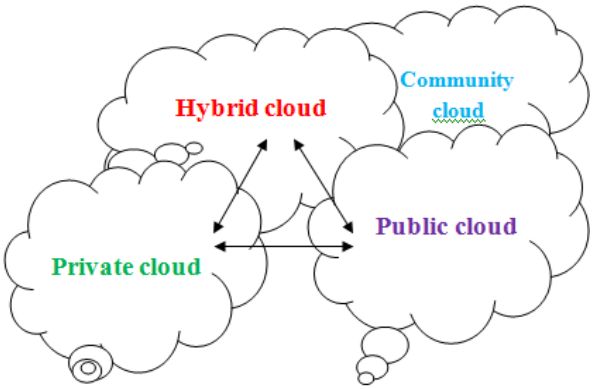

Fig 1: Types of cloud [16]

3. CLOUD COMPUTING SERVICE MODELS

Cloud computing providers deliver their IT services according to three fundamental models:

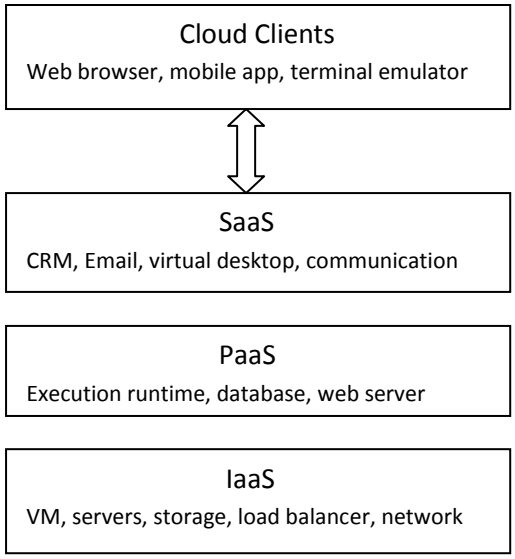

Fig 2: Cloud computing models [1]

3.1 Infrastructure as a Service (IaaS)

Includes servers, storage, virtual machines, load balancers and other core infrastructure stack IaaS cloud service provider offer this resource on-demand from their large pools installed in data centers [5]. Here we need to configure server through Network Administrator. For develop and Deploy software we need skilled IT developer team. The client is not responsible for managing the underlying infrastructure. Leading IaaS Service Provider are AmazonCloudFormation, Amazon EC2 Backspace, IBM and HP.

3.2 Platform as a Service (PaaS)

Adds development and programming models to IaaS. Includes databases, execution frameworks/runtimes, web servers and development tools.

Programmers can upload their application code to a particular platform without worrying about hardware infrastructure (IaaS). Here we need only programmer's team to develop and deploy software [5].

It acts as a mediator/channel between IaaS and SaaS models. Leading PaaS Service Providers are Google app Engine, Windows azure, force. comAWS Elastic Beanstalk, Cloud Foundry, Heroku, OrangeScape.

3.3 Software as a Service (SaaS)

In this model we only select software that is offered by service providers, and we can access the software from cloud through internet. It provides the facility to customize software as our requirement. The client is not responsible for managing the underlying infrastructure and platform on which application runs. The complete application offering in the cloud [5].Leading SaaS Service Provider areMicrosoftOffice365, Onlive, GTNexus, Marketo, and TradeCard, Salesforce CRM, Google Apps/Gmail/google drive/gtalk/GoogleCalendar, Microsoft "Live", Dropbox, and a lot more.

4. SLA (SERVICE LEVEL AGREEMENT)

Service level agreement (SLA) is a type of contract that must be signed between service provider and service customer before actual delivery of services. It defines functional and non-functional characteristic of a service including quality of services requirements and penalties in case of SLA violation. SLA document contains obligation and actions that will be taken in result of SLA violation [8].

Some of the SLA parameter such as Response time, Storage, Network bandwidth and memory which are used as metrics to determine the provisioning of requirement [7].The Service Level Objectives (SLO) defines the level of services that both parties agree on. Examples of SLA agreement:

Table 1. SLA agreement

SLA Parameter	**SLA Objective**
Response Time	< 10 ns
Storage	> 1TB
Memory	> 1GB
Network Bandwidth	> 1Gbps

CSP is a profit based company which may cheat with user by partially satisfying his/her request and make profit on them. This would lead to **SLA violation** and disrupt the communication between CSP and cloud user. For example: CSP may provide less memory/resources to cloud user while giving remaining memory to other cloud user to enlarge its overall profit with limited resources. Users don't have any idea about how much amount of memory he/she will get, as specified in SLA agreement.

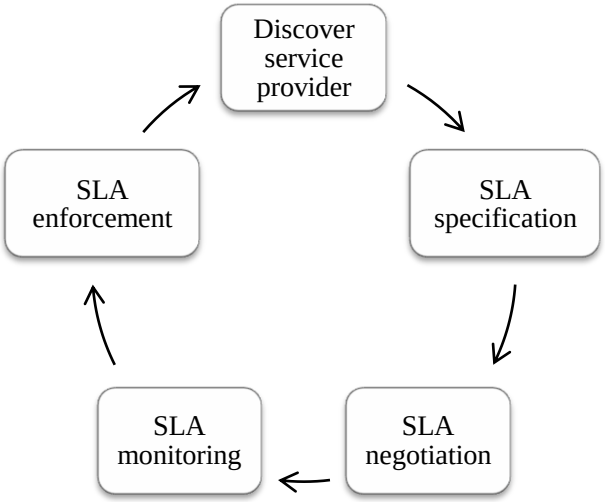

Fig 3: SLA Lifecycle

5. DIFFERENT SLA-VIOLATION REDUCTION POLICIES

5.1 VM Provisioning Method to Improve the Profit and SLA Violation of Cloud Service Providers [9]

In cloud computing environment there are two types of request like on-demand and reserved. The amount of on-demand resources required can vary per user request but it is unchanged for reserved type of request. Thus for accomplishing any task in the cloud environment we should consider current workload and idle capacity of each cloud provider. Threshold based approach provides efficient resource provisioning by taking current workload and request types into account. This approach works in cloud federated environment, in which if user request exceed the resource limit of cloud provider then cloud provider outsource to other cloud provider resources for satisfying the request. The basic Threshold model if the Load is less than threshold 1 then both type of request (reserved and on-demand) are accepted. If it is in between threshold1 and threshold 2 then only on-demand type request is accepted and if load is above the threshold 2 then no request is accepted.

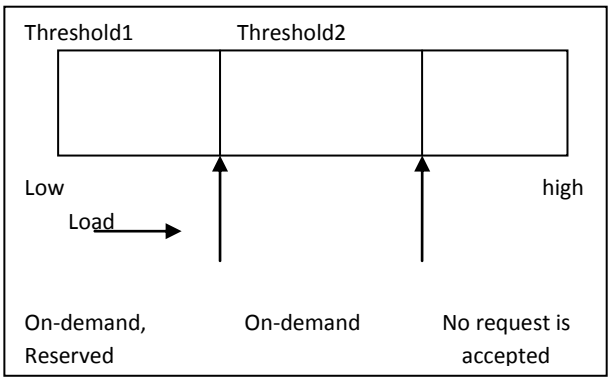

Fig 4: Threshold Values of Load

The proposed model uses the double threshold values for different types of request. If the request is of type on-demand and load is less than threshold 2 then VM is created otherwise it contact to other cloud coordinator. And if the load less than threshold 1 then both type of request (reserved and on-demand) are acceptedotherwise request of reserved type are rejected. This algorithm provides better resource provisioning for both type of request and better SLA based services to its users, which significantly reduces SLA violation. But it complicates the process of resource provisioning among multiple data center.

5.2 An Energy-Efficient Approach for Virtual Machine Placement in Cloud Based Data Centers [10]

Heavily loaded data centers in cloud computing environment consume huge amount of energy this would led to high carbon emission which is harmful for atmosphere. It has been estimated that in 2010 the energy consumed by the Information and communication technology (ICT) is 100 billion KWh and this had generated about 4% of the total global CO_2 emissions. Hence, energy-efficient techniques are required to minimize the negative impacts of Cloud computing on the environment. This paper proposes an energy efficient technique based on Minimum Correlation Coefficient (MCC) method and using fuzzy Analytic Hierarchy Process (AHP) for VM placement in cloud computing virtualized data centers. The proposed approach can make a suitable trade-off between SLA violation reduction and low energy consumption and also solves the problem of energy efficient VM placement, in reasonable time.*VM PLACEMENT USING PABFD-MCC METHOD:*For finding suitable host for VM placement, this gives proper balance between power consumption and SLA violation reduction in data centers. It uses two criteria for determining appropriate host according to our objectives:

1. Select a host that consumes less power after VM allocation.
2. Correlation coefficient between migrant VM and VMs running on target host (CC), in order to minimizes SLA violation.

Here Multiple Correlation coefficient is used in multiple regression analysis to assess the quality of the prediction of dependent variable. Minimum Correlation Coefficient (MCC) determines i^{th} observed value for dependent variable Y.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots\ldots \beta_p x_{pi} + e_i \quad \ldots (1)$$

Where β_0 is the constant value, β_1 to β_p is the coefficient of each independent value and e_i is error term. This method uses two matrixes X and Y every column in X represents a VM running on a target host, and we consider m current CPU utilization for each VM. X is m×(n+1) matrix. Y represents a migrant VM, and contains m current CPU utilization.
The correlation coefficient of CPU utilization between VMs which are going to migrate and other VMs which are running on the target host ($R^2_{y,1\ldots\ldots p}$) is calculated by

$$R^2_{y,1\ldots\ldots p} = \left(\frac{\sum_{i=1}^{n} (y_i - m_y)^2 (\hat{y}_i - m_{\hat{y}})^2}{\sqrt{\sum_{i=1}^{n} (y_i - m_y)^2 \sum_{i=1}^{n} (\hat{y}_i - m_{\hat{y}})^2}} \right)^2 \quad \ldots (2)$$

Fuzzy AHP method makes balance between SLA violation and Energy consumption of cloud datacenter.

5.3 SLA-based Virtual Machine Management for Heterogeneous Workloads in a Cloud Datacenter [11]

Cloud accommodates virtualized computer resources which host a heterogeneous application such as Non-interactive (batch jobs) and Interactive (transactional jobs), user-facing applications. Although virtual machine provides facility to run heterogeneous application concurrently, inspite of that cloud service provider unable to fulfill the complete SLA requirement. Example Amazon EC2 guarantees on availability of resources, not on VM Performance.This paper proposes Mixed Workload Aware Policy (MWAP) that incorporates an admission control and SLA enforcement mechanism that not only accelerates the resource utilization but also met the QoS requirement as specified in the SLA. This policy makes dynamic placement decisions to respond to changes in transactional type work load along with different SLA penalty parameters. To schedule batch jobs, this policy predicts the future resource availability and schedules jobs by stealing CPU cycles, which are under-utilized by transactional application for resource provisioning.

Admission Control makes the decision whether the requested VM (for an application) can be allocated based on present and future resource availability and the QoS requirements can be met with a given number of available resources.SLA Enforcement is handled by SLA Manager, which monitors the resource demand of different type of application based on QoS requirement as stated in SLA.Sla manager consider different types of SLAs along with their penalties for different types of application to avoid unnecessary penalty to cloud service provider. By intensive study it has been analyzed that MWAP is able to provision different workload and QoS requirement and also reduces the number of servers utilized by 60% over other strategies like vm consolidation and migration with negligible SLA violation, this would enhance the resource utilization and it can easily implemented in real cloud environment.

5.4 Response Time based Load Balancing in Cloud Computing [12]

In cloud computing one of the challenging issuesis load balancing. It needs to be maintained for efficiently utilizing the resources and satisfying the SLA requirement of customer request. This paper proposes a preventive approach that considers response time as primary factor for calculating the threshold value. It is dynamic in nature and also eliminates the negotiation cost between Load balancer and Virtual machine. Load balancer will choose a required VM if the combination of Average Response Time and Predicted Average Response time is less than current threshold time. This is Preventive algorithm, which scheduled the newly arrived request such that Load unbalancing will not occur. Algorithm has been divided in to three major parts:

a) Threshold Adjustment Module: This module work in two modes such as increment mode or decrement mode. In Increment mode, threshold value increases if none of the VM satisfies threshold value. And in decrement mode threshold get down, if the sum of highest Average Response time of all VM and half its value exceed a certain limit.

b) Average and Predicted Average Response time: They calculate the average response time and the predicted average response time for further updation in threshold time.

$$The\ average\ resopnce\ time\ of\ the\ VM = TT/N \quad \ldots (3)$$

TT = Total Time for Average Response Time Calculation.
N = total number of request encountered.

$$Predicted\ Time = newEX * PercentageAverage\ /100 \quad \ldots (4)$$

$$newEX = Responce\ Time\ of\ the\ newly\ arrived\ request.$$

$$Predicted\ Average\ Responce\ Time\ of\ This\ VM = ((average\ responce\ time\ of\ the\ VM) * N + Predicted\ Time\ )/(N+1) \quad \ldots (5)$$

c) Service allocation module: The new request to the particular VMs is assigned by this module on the basis of threshold value. Starting threshold value must be equal to max response time of each services on all VMs.

5.5 Energy-Saving Virtual Machine Placement in Cloud Data Centers [13]

Cloud is a grid of multiple service providers which have its own customers. Each customer have huge amount of data to be placed on cloud. For processing these huge amount of data, data center will consume huge energy and make higher outlays in cloud computing. According to statistics, each data center in world consumes as much energy as 250,000 households on average. Due to this overall cost of operating data center increases. To optimize energy consumption we need to consider both aspects such as physical resource (CPU, Memory, Storage, and Bandwidth) utilization as well as network resource utilization. This paper proposes VM Placement Algorithm to solve the multi-objective optimization problem. It is also a combination of network resource utilization and server resource utilization.

i) Optimization of Network resources: It uses network communication traffic to describe network power consumption Traffic matrix $A=(a_{i,j})_{N\times N}$, communication cost matrix $B=(b_{m,p})_{M\times M}$, $a_{i,j}$ is the traffic between VM i and VM *j*; $b_{m,p}$ represents the communication cost between PM m and PM p The objective function can be formally expressed as:

$$\min cost_{net} = \sum_{i,j=1}^{N} a_{i,j} b_{m,p} \quad \ldots (6)$$

ii) Optimization of Server Resources: It is used to minimize number of active PMs which minimizes the energy consumption. We define $X_{i,m}$as a binary variable, expressed as:

$$X_{i,m} = \begin{cases} 1 & if\ VM_i\ on\ PM_m \\ 0 & else \end{cases} \quad \ldots (7)$$

VM placement be formalized as follows:

$$\min cost_{ser} = \sum_{m=1}^{M} Y_m \quad \ldots (8)$$

$$\min cost_{net} + \min cost_{ser} \quad \ldots (9)$$

VM Placement Algorithm involves two steps:

Firstly, hierarchical clustering based on minimum cut algorithm enables certain VMs to cluster together to finally minimize the total network traffic.

Secondly, Best Fit (BF)is applied to optimize the PM resources and reduce PM's energy consumption.

Table 2. Comparison of Different SLA-Violation Reduction Policies

Technique	Method	Parameter	Outcome
VM Provisioning Method to Improve the Profit and SLA Violation of Cloud ServiceProviders	Threshold Based VM Provisioning	Static threshold	Minimize SLA violation
An Energy-Efficient Approach for Virtual Machine Placement in Cloud Based Data Centers	MCC and Fuzzy AHP method	Predicted utilization of host	Reduces Energy consumption and SLA violation
SLA-based Virtual Machine Management for Heterogeneous Workloadsin a Cloud Datacenter	Admission control and SLA based scheduling algorithm	Service level agreement	Increase utilization of resource in heterogeneous environment
Response Time Based Load Balancing in CloudComputing	Threshold adjustment and service allocation algorithm	Response time	Decreases the response time of user request
Energy-Saving Virtual Machine Placement in Cloud Data Centers	VM Placement Algorithm	Server and Network resource	Maximize the profit and reduce Energy consumption

6. CONCLUSION

The VM provisioning techniques are of two types. First one is Static VM provisioning and another one is Dynamic VM provisioning. The Static VM provisioning won't consider the dynamic workload of VMs. Therefore it is not well suited to provide low energy consumption while keeping low level of SLA violation. Whereas Dynamic VM provisioning consider the runtime workload of VMs and provides the more efficient performance over the Static technique. In this paper we have discussed the various SLA-violation reduction techniques. These techniques minimize the cost for resource of SaaS providers.SLA Violation may incur unnecessary penalty to cloud service providerand also lead to underutilizationof resources. Future work is try to minimize the SLA Violation as much as possible with low energy consumption.

7. REFERENCES

[1] Rahul Bhoyar, Prof. Nitin Chopde, "Cloud Computing: Service models,Types,Database and issues",International Journal of Advanced Research in Computer Science and Software Engineering (ijarcsse),Volume 3, Issue 3, March 2013.

[2] Andy Oram, George Rees, O Reilly Media,Sebastopol,"Cloud Application Architectures", Copyright © 2009.

[3] Mohiuddin Ahmed, Abu Sina Md. Raju Chowdhury, Mustaq Ahmed, Md. Mahmudul Hasan Rafee, "An Advanced Survey on Cloud Computing and State-of-the-art Research Issues",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[4] ShyamPatidar, DheerajRane, Pritesh Jain, "A Survey Paper on Cloud Computing", Second International Conference on Advanced Computing & Communication Technologies, © IEEE 2012.

[5] Eeraj Jan Qaisar, "Introduction to Cloud Computing for Developers," In IEEE ©2012.

[6] Wikipedia: http://en.wikipedia.org/wiki

[7] Lin Ye, Hongli Zhang, JiantaoShi Xiaojiang Du, "Verifying Cloud Service Level Agreement",© IEEEGlobecom 2012.

[8] Mohammed Alhamad, Tharam Dillon, and Elizabeth Chang, "A Survey on SLA and Performance Measurement in Cloud Computing",© Springer-Verlag Berlin Heidelberg 2011.

[9] Komal Singh Patel and A. K. Sarje, "VM Provisioning Method to Improve the Profit and SLA Violation of Cloud Service Providers", IEEE International Conference, Cloud Computing in Emerging Markets (CCEM) 11-12 Oct. 2012.

[10] NeginKord and Hassan Haghihi, "An Energy-Efficient Approach for Virtual Machine Placement in Cloud Based Data Centers", 5th Conference on Information and Knowledge Technology (IKT) 2013.

[11] Saurabh Kumar Garg, Adel NadjaranToosi, Srinivasa K. Gopalaiyengar, RajkumarBuyya, "SLA-based Virtual Machine Management for Heterogeneous Workloads in a Cloud Datacenter", Journal of Network and Computer Applications 1 August 2014.

[12] Agraj Sharma, Sateesh K. Peddoju, "Response Time Based Load Balancing in Cloud Computing", International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), © IEEE 2014.

[13] Jiankang Dong, Hongbo Wang, Yangyang Li, Peng Zhang, and Shiduan Cheng, "Energy-Saving Virtual Machine Placement in Cloud Data Centers", 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing 2013.