Cost Aware Dynamic Rule based Auto-scaling of Infrastructure as a Service in Cloud Environment

M. Kriushanth
Research Scholar in Computer Science,
St. Joseph's College (Autonomous) Tiruchirappalli,
Tamil Nadu, India.

L. Arockiam, Ph.D.
Associate Professor in Computer Science,
St. Joseph's College (Autonomous),
Tiruchirappalli, Tamil Nadu, India.

ABSTRACT
Cloud computing is one of the fastest growing technology. Pay-as-you-go model attracts the customer to utilize the large amount of cloud services in very low cost. Scalability and virtualization plays a vital role to achieve this goal. Scalability is the ability to find the number of users and to provide the service accordingly. Scaling can be divided into two, namely Auto-scaling or dynamic scaling and manual scaling. Auto-scaling doing great job to reduce the manual process. Scaling definitely reduces the service and operational cost, badly configured scaling sometimes increases the cost also. In such case there are chances for Service Level Agreement (SLA) violations and poor Quality of service (QoS). The perfect scaling should increase the profit for the Cloud Service Provider (CSP) and reduces the service cost, should not affect the QoS and SLA violations. In this paper, a dynamic rule based auto-scaling mechanism is proposed to reduce the cost of the VM instances.

Keywords
Cloud Computing, Auto-scaling, Virtualization, Quality of Service and Service Level Agreements.

1. INTRODUCTION
Cloud has become an attractive platform, it offers many on-demand computing power. As of today, it reached the commercial success because of the pay-as-you-go model. Scalability and virtualization play a vital role which helps in achieving the commercial success. Scalability is the ability to find the number of users and to provide the services accordingly. Scaling can be divided into two, namely Auto-scaling or dynamic scaling and manual scaling. Dynamic scalability helps users to scale up and scale down, scale in and out the computing resources. To reduce the administrators burden and to automate the process, cloud computing has the mechanism called auto-scaling [1]. Many cloud providers offer multiple types of Virtual machines (VM) with different capabilities and price. For example, Amazon offers VM instance types such as standard instances, high memory and high CPU instances [30]. Google provides the same like amazon with an added instance called shared core [2]. Cloud services are charged hourly basis, the rest of the hour is charged full hour by standard CSP's. For example, cloud user is using an instance for 1 hour and 25 minutes, the rest half of the service is considered next full hour and have to pay accordingly. To avoid this, google has launched a new pricing model for all VM types are charged minimum of 10 minutes. For example, if a user runs a instance for 2 minutes, user will be billed for 10 minutes of usage. After 10 minutes, instances are charged in 1 minute increments, rounded up to the nearest minute. For example, an instance that lives for 11.25 minutes will be charged for 12 minutes of usage [2]. Scaling can be done by proactive and reactive scaling, either horizontal or vertical scaling. The VM generally be acquired at any time, but could take several minutes to be ready [3]. The reactive scaling takes to instantiate a new VM takes 15 minutes to be ready [24]. Many works related to auto-scaling are proactive. Researchers try to predict the workload to avoid the over provisioning. The under provisioned mechanism will raise service cost for the user, overprovisioned mechanism leads to the profit loss for CSP and definitely will affect the QoS and SLA violations. This paper explores the dynamic rule based auto-scaling mechanism.

The rest of the paper is categorized and organized as follows. Section 2 provides the overview of the related work. Section 3 describes the motivation of the work. Section 4 enlists the problems occur in the area of concern. Section 5 elaborates the dynamic rule based auto-scaling mechanism and finally, section 6 concludes the paper and proposes the future work.

2. RELATED WORK
The dynamic resource provisioning in virtualized environment (e.g., [4][5][6][7][8][9][10][11]) projects attempt to achieve application performance goals through dynamic resource provisioning in virtualized environments. It mainly focuses on the use of control theory to tune up the application performance in a fine-grained manner. Several reactive and proactive workload prediction models were also proposed and evaluated. More recently, people have extended the idea into the cloud environment, such as [12][13][14]. These projects try to handle the workload by acquiring VM instances in the cloud when the computing capacity is insufficient.

The second category is cost-efficiency in the cloud. Several papers [15][16][17][18][19] discussed the resource allocation and instance consolidation strategies for cloud data centers. In general, the goal is to maximize cloud providers' profit and to maintain service level agreements (SLAs).

In cloud auto-scaling, most cloud providers offer APIs to allow users to control their cloud VM instance numbers programmatically, thus facilitating user-defined auto-scaling. AWS [1] and some third-party cloud management services such as RightScale [18], enStratus [19], Scalr [20], etc. offer schedule-based and rule-based auto-scaling mechanisms. Schedule-based auto-scaling mechanisms allow users to add and remove capacity at a given time, such as "run 10 instances between 8AM to 6PM each day and 2 instances all the other time". Rule-based mechanisms allow users to define simple triggers by specifying an instance scaling thresholds and actions, such as "add (remove) 2 instances when the average CPU utilization is above 70% (below 20%) for 5 minutes".

Yang et al. [24] introduced a cost-aware auto-scaling approach using the workload in service clouds. The ultimate aim of their experiment is to reduce the SLA violation and to keep the scaling cost low and also to improve the QoS.

The cost of the service cloud will be less if we use less virtual resources from the cloud provider, but performance will be affected when the peak load occurs. When resources are used appropriately, leasing more virtual resources may leads to a performance improvement if the cloud but also bears a higher cost. To maintain the SLA and QoS while still keeping costs low is challenging task due to frequent variation of workload. CSP can use vertical or horizontal scaling for the various categories of instances and cost also differ in both. They proposed the workload prediction is best method to keep service cost low, for prediction, linear regression analysis and a second order Auto Regressive Moving Average method filter (ARMA) are used. They compared pre-scaling, real-time and auto-scaling and finally proposed a new framework to reduce the operational cost.

Zhang et al. [26] had made an extensive study on the research challenges and issues in cloud computing. Various aspects of cloud computing were discussed such as automated service provisioning, virtual machine migration, server consolidation, energy management, traffic management and analytics, data security, software frameworks, storage technologies and data management and novel cloud architecture. They have considered automated service provisioning and server consolidation in their article. The objective of a service provider is to allocate and de-allocate resources from the cloud to satisfy its SLA, while minimizing its operational cost. They have concentrated on dynamic service provisioning and typically involved approaches. Constructing an application performance model that predicts the number of application instances required to handle the demand at each particular level, in order to satisfy the QoS requirements. Using the performance model, the future demands are periodically predicted and the resource requirements are determined. The application performance model could be constructed using various techniques including Queuing theory, Control theory and Statistical machine learning.

The server consolidation concept is used to maximize the resource utilization while minimizing energy consumption. Multiple underutilized servers are merged into a single server, so that the remaining server can be set to an energy saving state. The cost of powering and cooling the data centers takes 53% of the total operational expenditure.

Dynamic scalability enables users to quickly scale up or down underlying infrastructure. Several challenges arise when considering computing instances such as non-deterministic acquisition time, multiple VM instance types, unique billing models and user budget constraints. Deadline and budget constraint for cloud infrastructure to accommodate changing workload based on application level performance metrics job deadline. [21] have used windows Azure to test the experiment which takes 10 minutes to start an instance and shutting time quite stable around 2-3 minutes. VM startup delay plays an important role in cloud Auto-Scaling mechanism. In their experiment, VM instances were billed by hours; fraction consumption of an instance hour was also counted as a full hour. Ming et al. [21] proposed an architecture to finish the job before the deadline to bring out the cost effectiveness. VM startup delay could not only affect the performance, but also dominates the utilization rate and the cost for short deadline.

The goal of cloud computing is to allocate the resources that are needed for the customers and charge accordingly. The service providers like Amazon EC2 [30] currently offers 11 VM instance types like a standard machine for most types of application, high CPU and high memory used to finish the job before stipulated time or the deadline. Ming et al. [22] accomplished the goal by dynamically allocating/de-allocating VMs as a scheduled task on most cost efficient instances. The evaluation of the experimental results have shown the total cost saving from 9.8% to 40.4%.

The auto-scaling techniques basically used to automate the scaling and to reduce the waiting time and cost. Tania et al. [23] proposed an auto-scaling techniques for elastic application in cloud environments. Auto-scaling can be done by proactive or reactive scaling, proactive is much more cost effective. The first approach is static threshold-based rules. When the CPU utilization has reached 70% for more than 5 minutes, it adds 2 instances. If it reaches 30%, it will reduce the instances. The performance metrics can be considered by request rate, CPU load or average response time.

3. MOTIVATION

From the above brief literature analysis, it is known that there are several methods proposed which addresses the auto-scaling issues. But, no mechanism fulfils the four parameters (Cost, Time, SLA and QoS). Some of the works have used the static threshold-based rule for providing VM instances which increases the service cost for the user. Some of the works violated the Service Level Agreements which obviously reduces the QoS. It is essential to provide an auto-scaling mechanism which satisfies all the QoS parameters. The forth coming section describes the proposed dynamic rule based auto-scaling mechanism for providing VM instances to the users which aims to achieve the reduced cost.

4. AREA OF CONCERNS

The following are the areas where the problems occur in the cloud environment.

4.1 Cloud Services and pricing

From CSP's perspective, cloud service cost included the operational cost, power, cooling system and floor space.

- The cloud service providers offer various types of VMs with different capabilities. Standard, High-memory and High-CPU offered by Amazon Web Services [3].
- Standard, High-memory, High-CPU and Shared core by Google [2].
- Extra small, small, Medium, Large and Extra-large by Microsoft [28].
- Type I – VII it varies in size and capability by Rackspace [29].

As per current scenario, VM instances are priced by hour. The partial-hour consumption is always rounded up to one hour. A minute based billing model also possible by some CSPs. The cloud instances can be acquired at any time. It may take some minute to be ready to be used. Such case user has to wait in a queue for the service, in such situation is called as waiting time. In real time scenario workload is heavy in daytime compared to night and week day and week end as well.

4.2 Workload

Workload can be segregated in three types, small medium and heavy. Expect heavy workload other types are manageable, in sudden peak situation, CSP keep vigil on the request and work load, It may change all the time. Reactive techniques are always time consuming also may mislay user satisfaction. Proactive techniques are always preferable to avoid such

complications. Most of the related work based on the prediction and workload forecasting [25].

4.3 Virtual Machine

Usually VM startup time in minutes, at the end user side time taken to get a service are included parameters like time of the day, VM instance type, OS, location of the data center and number of requested instance at a time [27].

4.4 SLA and QoS

Any one of the above scenario is affected, there are chances to violate the SLA definitely and cannot guarantee the QoS.

4.5 Threshold value

Threshold value (Upper Threshold & Lower Threshold) ratios can be any form 60% - 20%, 70%-30%, and 90%-10%. According to Tania et al. [31] threshold values are compared with three proactive methods based on time series analysis. Moving Average (MA), linear regression (LR) and Exponential smoothing (ES) are considered. Regarding proactive techniques, 60%-20% threshold configuration obtained the lowest number of SLA violations and VM cost is bit higher. When 90%-10% threshold configuration VM cost is affordable and SLA violations are high. Finally 70%-30% threshold value provides the better VM cost and medium level SLA violations.

Beyond the upper threshold value, Maximum utilization limit of the resource could not find in most of the research work.

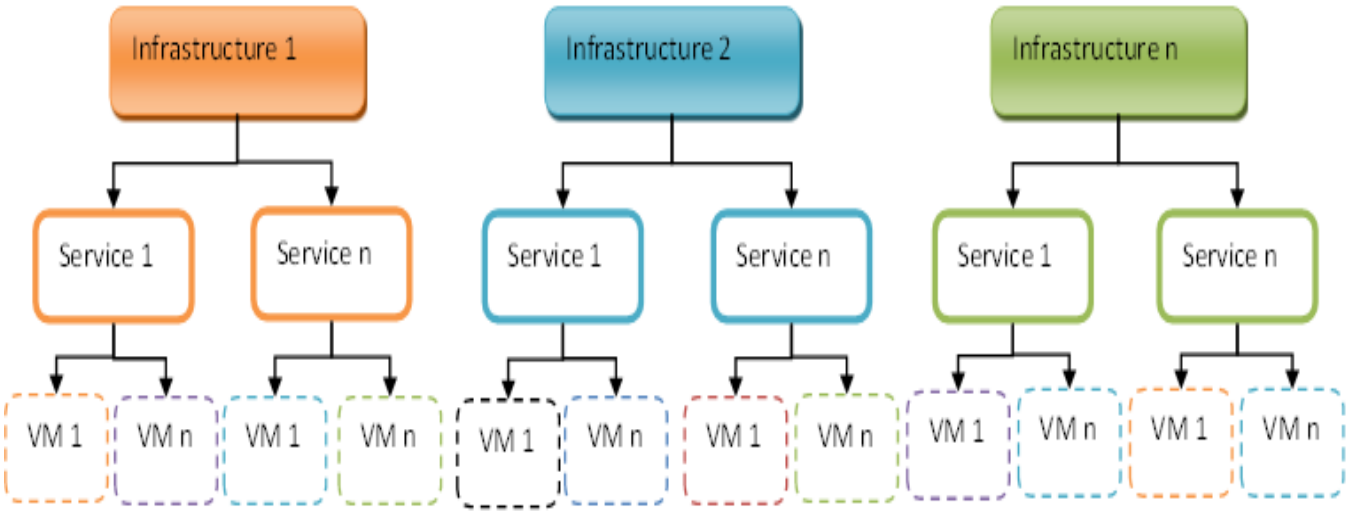

Figure 1. IaaS VM Clusters

5. PROPOSED WORK

The resources can be divided in to capability wise (small to High) and allocated to the users like in Fig 1. Consider a scenario, if the high, medium and standard category resources allocated in various clusters separately (Ex. High CPU is requests allocated in infrastructure 2 and Standard is allocated in Infrastructure 1). If any one of VM gets failure, the low category services may not cause serious problem, high end VM clusters encountered serious problem, it may affect the Cost, SLA and QoS. To avoid such situation, all types of incoming requests are allocated consecutively in all available infrastructures.

5.1 Rule based Auto-scaling Framework

A framework to implement the cost aware dynamic rule based auto-scaling approach is shown in Figure 2. This focuses on the management of the user request, rule engine and VM instances.

5.1.1 Admission controller (ADC):

User request is sent through the admission controller. It filters out the invalid login requests and allows only the authorized users to use the service.

5.1.2 Load balancer (LB):

Load balancing is a method for distributing workloads across multiple computing resources, such as computers, a computer cluster, network links, central processing units or disk drives. The load balancer collects the server status from the server cluster, then forwards request to the suitable VM according to the load balancing strategy.

5.1.3 Workload Analyzer (WA):

Workload analyzer collects the information from the server cluster's history log and predicts the workload. It analyses the workload for the next time. Next, the rule engine decides which rule is cost effective, according to the analyzed results.

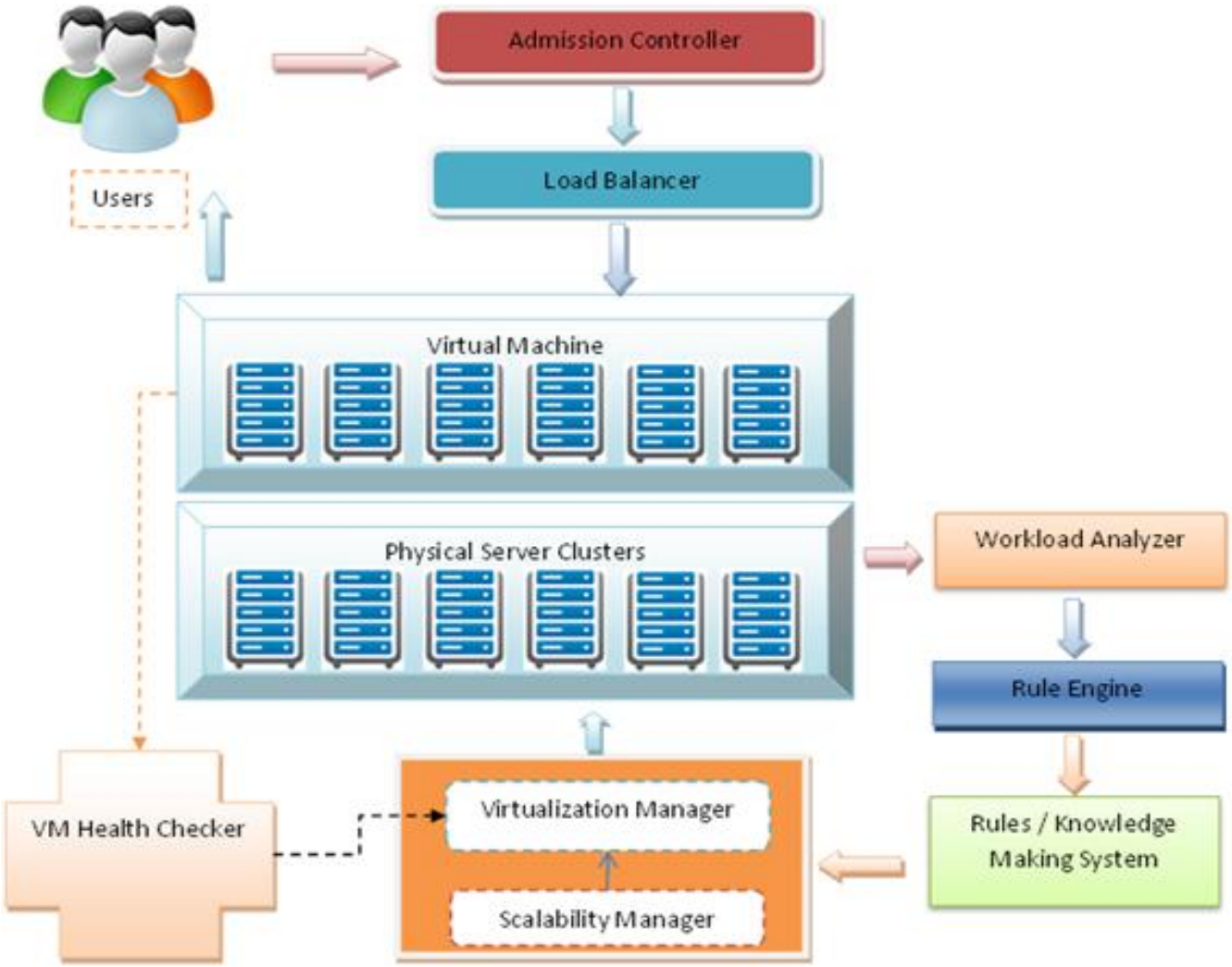

Figure 2 Rule Based Auto-Scaling

5.1.4 Rule engine (RE):

Rule engine generates the rule with analyzed request. If the workload reaches 70% threshold value add 2 instances and reaches below 30% remove the instances.

5.1.5 Rules / Knowledge making system (RMS):

According to the rule engine direction, rules making system creates some knowledge to manage the workload till the new VM is ready. If the requests reach 70%, a rule is applied to check the ability to handle the requests. If capable, the threshold value is increased to its capacity (e.g. 10%-20%). If the requests reach below 20% threshold value, remove the underutilized VM. Meanwhile, deal the request with available VM and direct the new request to the available or free VMs.

5.1.6 Scalability Manager (SM):

The scalability manager is the essential component in the scaling framework. The auto-scaling works here. It decides when and how to scale the service, as per knowledge making system.

5.1.7 Virtualization Manager (VM):

Virtualization manager deals with the virtual resources in the physical server clusters. It executes policies proposed by scalability manager.

5.1.8 VM Health checker (VMHC):

The virtual machine health checker keeps vigil on the running VMs. If any VM instance is found unhealthy or not running properly, it sends a message to the virtualization manager to replace the unhealthy VM instances.

5.2 Algorithm for Dynamic Rule-based Auto-scaling

```
1   Begin
2   Predict the initial work load and boot VMs;
3   if (CTV >= NT)
4       Check CPY
5       if (CPY==true)
6           Manage all the instances
7       else
8           Add 2 instances
9   else if (CTV > MT)
10      Manage all instances
11  else if (CTV > UT)
```

12 Send all requests to the new VM

13 else if (*CTV* < *LT*)

14 Remove VM

15 else

16 manage the request with available VM

17 End

Table 1. Components of the dynamic rule base auto-scaling algorithm

Component	Description
CTV	Current threshold value
NT	Normal threshold
MT	Medium threshold
CPY	Capability of VM
UT	Upper threshold
LT	Lower threshold

6. CONCLUSION AND FUTURE WORK

Scalability and virtualization plays a vital role to achieve this goal. Scalability is the ability to find the number of users and to provide the service accordingly. This paper investigated the problems that happens in auto-scaling in rule based approach for cost aware services. In this paper, the cost aware dynamic rule based auto-scaling approach is proposed to reduce the cost of the service. Our approach will reduce the cost of the service and SLA violations related to cost. Our future work will be reducing the waiting time of the user.

7. REFERENCES

[1] "Amazon Auto Scaling in Cloud Computing", http://aws.amazon.com/autoscaling/30.05.2013.

[2] https://developers.google.com/compute/pricing

[3] "Amazon EC2 pricing ", https://aws.amazon.com/ec2/pricing

[4] Z. Hill, J. Li, M. Mao, A. Ruiz-Alvarez and M. Humphrey, "Early Observations on the Performance of Windows Azure", Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. Chicago, Illinois, June 21, 2010.

[5] H. Lim, S. Babu, J. Chase, and S. Parekh, "Automated Control in Cloud Computing: Challenges and Opportunities", 1st Workshop on Automated Control for Datacenters and Clouds, June 2009.

[6] P. Padala, K. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, A. Merchant, and K. Salem. 2007. Adaptive Control of Virtualized Resources in Utility Computing Environments. EuroSys, 2007.

[7] B. Urgaonkar, P. Shenoy, A. Chandra, and P. Goyal, "Dynamic provisioning of multi-tier internet applications", 2nd International Conference on Autonomic Computing, Seattle, WA, USA, June 2005.

[8] Q. Zhang, L. Cherkasova, E. Smirni, "A Regression-Based Analytic Model for Dynamic Resource Provisioning of Multi-Tier Applications", Proceedings of the Fourth International Conference on Autonomic Computing, Jacksonville, Florida, USA, June 2007.

[9] A. Chandra, W. Gong and P. Shenoy, "Dynamic Resource Allocation for Shared data centers using online measurements", Proceedings of the 11th International Workshop on Quality of Service, 2003.

[10] J. Chase, D. Irwin, L. Grit, J. Moore, and S. Sprenkle, "Dynamic Virtual Clusters in A Grid Site Manager", Proceedings of the 12th High Performance Distributed Computing, Seattle, Washington, June 22-24, 2003.

[11] S. Park and M. Humphrey, "Feedback-Controlled Resource Sharing for Predictable eScience", Proceedings IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC08),Austin, Texas, 2008.

[12] S. Park and M. Humphrey, "Predictable High Performance Computing using Feedback Control and Admission Control", IEEE Transactions on Parallel and Distributed Systems (TPDS), Mar, 2010.

[13] P. Ruth, P. McGachey and D. Xu, "VioCluster: Virtualization for Dynamic Computational Domains", Cluster Computing, IEEE International, pages 1-10, Sep 2005.

[14] Y. Yazir, C. Matthews, R. Farahbod, S. Neville, "Dynamic Resource Allocation in Computing Clouds using Distributed Multiple Criteria Decision Analysis", 3rd International Conference on Cloud Computing, Miami, Florida, USA, 2010.

[15] M. Mazzucco, D. Dyachuk and R. Deters, "Maximizing Cloud Providers Revenues via Energy Aware Allocation Policies", 3rd International Conference on Cloud Computing, Miami, Florida, USA, 2010.

[16] I. Goiri, J. Guitart and J. Torres, "Characterizing Cloud Federation for Enhancing Providers' Profit", 3rd International Conference on Cloud Computing, Miami, Florida, USA, 2010.

[17] F. Chang, J. Ren and R. Viswanathan, "Optimal Resource Allocation in Clouds", 3rd International Conference on Cloud Computing, Miami, Florida, USA, 2010.

[18] E. Deelman, G. Singh, M. Livny, B Berriman, and J. Good, "The Cost of Doing Science on the Cloud: The montage example", Proceeding SC '08 Proceedings of the 2008 ACM/IEEE conference on Supercomputing. pp. 1-12. 2008.

[19] RightScale. http://rightscale.com

[20] enStratus. http://www.enstratus.com

[21] Scalr. https://www.scalr.net

[22] Ming Mao, Jie Li and Marty Humphery, "Cloud Auto-scaling with Deadline and Budget Constraints", Proceedings of 11th IEEE/ACM International Conference on Grid Computing, 2010, ISBN 978-1-4244-9349-4, pp 41-48.

[23] Ming Mao and Marty Humphery, "Auto-scaling to Minimize Cost and Meet Application Deadline in Cloud

Workflows", Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE 2011, E-ISBN 978-1-4503-0771-0, pp 1-12.

[24] Tania Lorido-Botran, Jose Miguel-Alonso and Jose A. Lozano,"Auto-scaling Techniques for Elastic Applications in Cloud Environments", Technical Report EHU-KAT-IK-09-12, Department of Computer Architecture and Technology, University of the Basque Country, September 5, 2012, pp1-44.

[25] Jingqi Yang, Chuanchang Liu, Yanlei Shang, Bo Cheng, Zexiang Mao, Chunhog Liu, Lisha Niu and Junliang Chen, "A cost-aware auto-scaling approach using the workload predection in service clouds", Springer 2013.

[26] Qi Zhang, Lu Cheng and Raouf Boutaba, "Cloud Computing: state-of-the-art and research challenges", Springer, April 2010, pp 7-18.

[27] Ming Mao and Marty Humphery, "A Performance Study on the VM Startup time in the Cloud", IEEE 5th International Conference on Cloud Computing, IEEE Computer Society, 2012, pp 423-430.

[28] "Microsoft instance types", http://msdn.microsoft.com/en-us/library/azure/dn197896.aspx/ 30.04.2013

[29] "Rackspace instance types", http://www.rackspace.com/cloud/servers/pricing/ 30.04.2013.

[30] AWS available instance types https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-types.html.

[31] Tania Lorido-Botran, JoseMiguel-Alonso and Jose A. Lozano," Comparison of Auto-scaling Techniques for Cloud Environment", In. Proc. of CDEI, 2013, ISBN: 978-84-695-8330-2.

8. AUTHOR'S PROFILES

Kriushanth. M is a full time Ph.D. research in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has years of teaching experience. He has attended many International and National Conferences, Seminar and Workshops. His area of research is Cloud Computing. He is presently working on Scalability issues in Cloud Computing. His areas of interest Computer Networks and Web Technologies.

Dr. Arockiam. L is working as Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 25 years of experience in teaching and 17 years of experience in research. He has published more than 187 research articles in the International / National Conferences and Journals. He has also presented 2 research articles in the Software Measurement European Forum in Rome. He has chaired many technical sessions and delivered invited talks in National and International Conferences. He has authored a book on "Success through Soft Skills". His research interests are: Big Data, Cloud Computing, Software Measurement, Cognitive Aspects in Programming, Data Mining and Mobile Networks. He has been awarded "Best Research Publications in Science" for 2010, 2011, & 2012, "Best Teacher Award" for 2012-13, 2013 -14 and ASDF Global Awards for "Best Academic Researcher" from ASDF, Pondicherry for the academic year 2012-13.