ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 9, November 2012

A Cost Effective Scaling Approach for Cloud Applications

Bhushan Lal Sahu, Rajesh Tiwari

Abstract— **Cloud computing is becoming an increasingly popular enterprise model in which computing resources are made available on-demand to the user as needed. In IT services the demand is not fixed and not predictable, hence there is need of a good capacity planning to better serve of cloud applications. Scalability and virtualization play the key role to capacity planning in cloud computing. Scalability is the best solution to increasing and maintaining application performance in cloud computing environments. There are two main methods to scaling the cloud environments- scale up (vertical scaling) and scale out (horizontal scaling) but for most of the cloud environment scale out is preferred. In this paper we discussed pros and cons of both the method in cloud environment and proposed a cost effective scaling approach for the applications in cloud environment based on resource utilization where scale-up firstly prefer and then Scale-Out.**

Index Terms— **Cloud computing; Hybrid Scaling; IaaS; PaaS; Scale-Out; Scale-Up; virtualization;**

I. INTRODUCTION

Cloud computing [1],[2],[3],[4] may be defined as an abstraction of services from infrastructures (i.e. hardware), platforms and applications (i.e. software) by virtualization of resources. The different forms of cloud services are IaaS, PaaS, and SaaS, which is Infrastructure-as-a-Service, Platform-as-a-Service, and Software-as-a-Service, respectively. Most IaaS and PaaS providers provide more or less detailed information on how scaling is done in their products, there are almost no details available from SaaS providers. On the one hand, we can argue that SaaS providers do not reveal how scaling is done in their products because this information is not important for users. What is important for users is that their applications do scale automatically. Hence in this paper we think about scaling on IaaS and PaaS levels only. Fig.1 [5] provides an overview of the mechanisms handy to accomplish the goal of whole application scalability.

Cloud Computing offers simple virtual hardware infrastructure for instance VMs and networks are usually termed Infrastructure as a service Clouds (IaaS) [6],[7]. A different abstraction level is done by Platform as a Service (PaaS) clouds. PaaS clouds provide a container-like environment where users deploy their applications as software components [8]. PaaS Clouds are becoming important elements for building applications in a faster manner even though they are less flexible than IaaS Clouds [6],[9] and various important IT players such as Google and Microsoft have developed PaaS clouds systems such as Google Apps Engine, Microsoft Azure, Jelastic Clouds.

Scalability is the best solution to increasing and maintaining application performance in cloud computing environments. Scalability or ability to expand and add resources dynamically is the success of many enterprises currently involved in doing business on the Web and in providing information that may suddenly become heavily demanded [10]. The two main scaling method used in cloud environment horizontal scaling (i.e. adding new server replicas and distribute load among all existing replicas using load balancer) and vertical scaling (on-the-fly changing of the assigned resources to an already running instance, for instance, adding more physical CPU to a running virtual machine (VM)) [5].

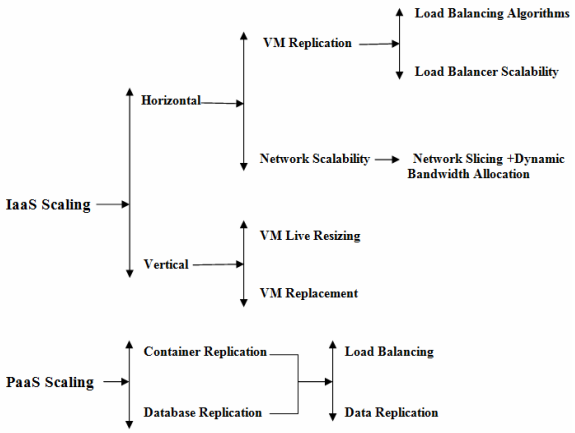

Fig. 1. Summary of the Available Mechanisms for Holistic Application Scalability

Virtualization, in computing, is the formation of a virtual (rather than actual) version of something, such as a hardware platform, operating system, a storage device or network resources. Virtualization can optimize resource sharing among applications hosted in different virtual machines to better meet their resource needs. As a result increasingly computing can be conducted in shared resource pools that act as private and public clouds. The huge processing power of Cloud Computing is made possible though distributed, large-scale computing clusters, often in concert with server virtualization software, like VMware ESX Server and Xen, and parallel processing. This network of servers and connections is collectively known as the Cloud. The model of Cloud Computing has evolved from following concepts of utility computing, autonomic computing, grid computing, and software as a service (SaaS) [11]. Cloud Computing makes different available resources on-demand. The

Manuscript received Nov 14, 2012.
***Bhushan Lal Sahu**, Department of Computer Science and Engineering, SSCET, Bhilai, CG, India.*
***Rajesh Tiwari**, Department of Computer Science and Engineering, SSCET , Bhilai, CG, India.*

on-demand nature of Cloud Computing combined with the pay-as-you-go model that means as the application demand grows, so can the resources required to service that demand [10].

In this paper, we will present a cost effective scaling approach which combining both horizontal and vertical scaling based on workload in virtualized cloud environment for better resource utilization and cost control. We will first describe Cloud Computing and the various scaling point for scaling cloud infrastructure. We then discuss the capabilities of the Cloud and its use of virtualization technologies. We will then present our novel architecture design of a scaling scenario with an on-line cloud application.

The rest of this paper is organized as follows: Section II describes the cloud computing architecture and virtualization. Section III presents the available scaling approaches to cloud infrastructure and also provide view of scaling points. The proposed scaling approach will be discussed in Section IV. Finally, Section V concludes the paper.

II. Cloud Architectures and Virtualization Technology

Cloud Computing provides the ability to add capacity as needed, typically with very small lead times.In this section we discuss about the cloud service architecture and key delivery technology of virtualization in cloud computing environment.

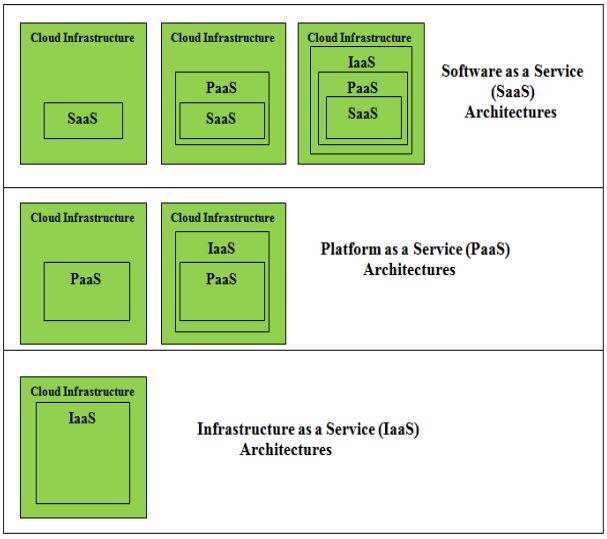

Fig. 2. Cloud computing Services

A. Cloud Architectures

From Fig. 2, it is possible to identify three Cloud Service Models, namely IaaS, PaaS and SaaS.

1) Infrastructure as a service (IaaS): IaaS is the delivery of computer infrastructure (typically a platform virtualization environment) as a service. IaaS leverages significant technology, services, and data center investments to deliver IT as a service to customers. In IaaS the user should opt from virtual computers, cloud storage, network infrastructure components such as firewalls and configuration services. Usage fees are calculated per CPU hour, data GB stored per hour, network bandwidth consumed, network infrastructure used per hour, value added services used, e.g., monitoring, auto-scaling etc.

2) Platform as a Service: PaaS offerings facilitate deployment of applications without the cost and complexity of buying as well as managing the underlying hardware and software and provisioning hosting capabilities, providing all of the facilities required to support the complete life-cycle of building and delivering web applications and services entirely available from the Internet [12]. PaaS is a platform where software can be developed, tested and deployed. It means the entire life cycle of software can be operated on a PaaS. Well known examples of PaaS include Google Apps Engine (GAE), Microsoft Azure, IBM SmartCloud, Amazon EC2, salesforce.com and jelastic.com.

3) *Software as a Service:* Software-as-a-Service (SaaS) is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet. SaaS is often associated with a pay-as-you-go subscription licensing model. In SaaS, the consumer is free of any worries and hassles related to the service. The Service Provider has very high administrative control on the application and is responsible for update, deployment, maintenance and security. For example, Gmail is a SaaS where Google is the provider and we are consumers. We have very limited administrative and user level control over it, although there is a limited range of actions, such as enabling priority inbox, signatures, undo send mail, etc, that the consumer can initiate through settings.

B. Virtualization Technology

Cloud computing has become popular because it provides a way to increase capacity or add capabilities on the fly without investing in new infrastructure, training new personnel, or licensing new software and virtualization technology [13] play the key delivery technology. Through Virtualization cloud computing removes the dependencies between software and the hardware that runs it. Virtualization allows multiple Operating Systems to share a single physical interface, to maximize the utilization of computer system resources, such as I/O devices. A virtual machine is a tightly isolated software container that can run its own operating systems and applications as if it were a physical computer. A virtual machine behaves exactly like a physical computer and contains it own virtual CPU, RAM hard disk and network interface card (NIC). However, a virtual machine is composed entirely of software and contains no hardware components at all. As a result, virtual machines offer a number of distinct advantages over physical hardware. An additional software layer is introduced to provide the illusion of Virtual Machines (VMs), called VMM (Virtual Machine Monitor) or hypervisor [14], on top of which each OS assumes owning resources exclusively. There are mainly two approaches to enable virtualization: Full virtualization and para-virtualization [13],[15]. The current leading virtualization and software providers include VMware, Xen and so on.

III. CLOUD SCALABILITY

In this section we discuss about the cloud scaling points and available scaling method of cloud.

A. Scaling Points

Scaling is the ability to increase or decrease compute capacity either by launching additional servers or changing server sizes. The on-demand nature of cloud computing means as application demand grows, resources use to service that demand and for this it is necessary to understand how a cloud application consumes resources, how it behaves under high load, and what happens for the potential scaling points of the entire system in order to maintain the desirable performance of the cloud application. One method to understand how the application performs under increased customer load is to explore where the scaling points are, and to use capacity planning to measure load on these points to add more resources when they are needed [10]. Depending upon the application every server has stress point to modify the resources. The typical scaling point includes [16]:

- Storage capacity (GB)
- Server processing power (CPU cycles) & RAM capacity (GB)
- Network bandwidth (Gbps)
- Database transactions per second (TPS)
- Storage input/output operations per second (IOPS)

Once a scaling point is selected for use to scale the application, samplings of the scaling indicator are collected in real-time, and statistics is calculated periodically. Based on the historical trends and predictions derived from the statistics of the scaling indicator, scaling rules can be defined to scale up or down the amount of application instances and capacity per instances [10].

B. Cloud Scaling

The cloud computing empowers to increase or decrease computing resources to meet load requirements. We can scale our capacity both manually (by executing a command on a command line or through a web interface) and dynamically (through predefined changes in capacity or through software that automatically adjusts capacity to meet actual demand). The ability to manually adjust capacity is a huge advantage over traditional computing. But the real power of scaling in the cloud lies in dynamic scaling also refers as auto scaling [17],[18]. The obvious benefit of cloud scaling is that you pay only for the resources you use. The noncloud approach is to buy infrastructure for peak capacity, waste resources, and pray your capacity planning was spot on. The downside of cloud scaling, however, is that it can become a crutch that lazy system architects use to avoid capacity planning. There are two main methods of scaling a cloud compute platform, namely scaling up (vertical scaling) and scaling out (horizontal scaling).

1) *Scaling Up:* Scaling up means ability to scale the size of a server either by resizing the server or by replacing that server to bigger one. Scaling up typically involves the addition of CPUs or memory to a single computer. Such vertical scaling of existing systems also enables them to use virtualization technology more effectively, as it provides more resources for the hosted set of operating system and application modules to share. Depending on application's architecture, cost, and bandwidth requirements, it might be more beneficial to scale vertically instead of horizontally. For example, instead of managing 6 smaller servers, 2 larger servers can provide an equivalent amount of performance at a more cost-effective rate. Vertical scaling is also common for growing the size of your database over time. In hardware terms; this includes adding processing power and memory. In software terms, scaling up may include optimizing algorithms and code.

2) *Scaling Out:* Scaling out requires the addition of more machines or devices to the computing platform to handle the increased demand. For example, as application grows and the number of user requests increases, scale-up horizontally by launching and provisioning additional server resources to serve the application. Conversely, when those resources are no longer needed, scale-down and automatically terminate underutilized server resources. In this method of cloud scaling the server/machine size is fixed and we can add to the computing platform by replication of same server to handle the increase demand.

Both the methods, Scaling Up and Scaling Out, have pros and cons that are important to understand for better resource management and cost effective approach for cloud applications. Table-I summarizes the differences between these two [16].

Table I. Difference between Horizontal and Vertical Scaling Method

Scaling Methods	How Scales	Scalability	Complexity	Capital Costs
Horizontal	More	Massive	High	Low
Vertical	Bigger	Moore's law-based 1	Low	High

Traditionally, most businesses have best served by using vertical scaling (Up) methods as long as possible and then scaling individual parts of application horizontally (Out) but in Cloud environment the scenario is changed and most businesses firstly served by using horizontally because the most common operating systems do not support on-the-fly (without rebooting) changes on the available CPU or memory to support this "vertical scaling" [5]. Vertical scaling typically involves making significant changes to a server's core configuration. Therefore, it's better to perform such changes manually and when try to set up scalable server arrays for (horizontal) auto scaling purposes, and then cannot change an existing server's configuration. Rather, can simply add or subtract compute resources based on changing demand metrics just like Amazon Auto scaling [19]. Amazon is one of the major players in providing cloud computing services like IaaS and PaaS. Amazon EC2 one of the main product of Amazon that allows users to rent virtual machines means IaaS cloud service. For handling most sudden, temporary peaks in application demand vertical scaling is the most apt and horizontal scale neither fully utilized resources nor entirely solve performance issues, especially those related to disk and network I/O [20]. But the problem is vertical scaling is limited by underlying hardware as well as most common operating systems do not support on-the-fly

(without rebooting) changes on the available CPU or memory [5]. Today there are many cloud service provider which provide vertical scaling of cloud application. Jelastic Cloud is one of them which provide vertical scaling for java applications automatically as Platform as a Service [21]. In Jelastic when application needs more CPU or RAM then it provides the resources automatically, and when these are no longer in use then automatically reduce them to save costs. In Jelastic users/customers do not require to select the machine size and pay for it. Instead, Jelastic dynamically allocates resources instantaneously scaling application servers up and down, and making sure that they have the resources that they need. These charge users for the actual RAM and CPU usage rather than for predefined machine size. Users benefit because the costs automatically go down when applications are off or not being used. It measures resources are being consumed in cloudstep (Additional RAM and CPU). Another example is GoGrid that provides dynamically scaling of cloud server's RAM on-demand i.e. dynamically increase as well as decrease the amount of RAM allocated to a particular cloud server instance.

IV. Proposed Scaling Approach

In the following section, we will present the novel architecture design, and the cost effective scaling approach of application installed in virtual machine on a cloud.

As mentioned previously, most of the operating system does not support vertical scaling without rebooting but there are many cloud provider that provide automatic resource provisioning using vertical scaling [16],[21]. We are proposing a cost effective scaling approach for cloud applications which is combination of both scale up and scale out as shown in Fig 3. The load balancer accepts incoming requests from clients and efficiently distributes requests to perform state-checking functionality to monitor the utilization of back-end servers. This is load based scaling approach i.e. as application traffic grows, the cloud platform automatically adds the CPU and RAM the application requests to handle the load that means when the load goes up, it gets the new resources it needs automatically and instantly, without waiting for the platform to re-deploy applications or re-configure the complete environment. If traffic decreases, instantaneously reduce the resources again. In this way application can handle any load/traffic without architectural changes or manual adjustments, and you pay only for resources you actually need.

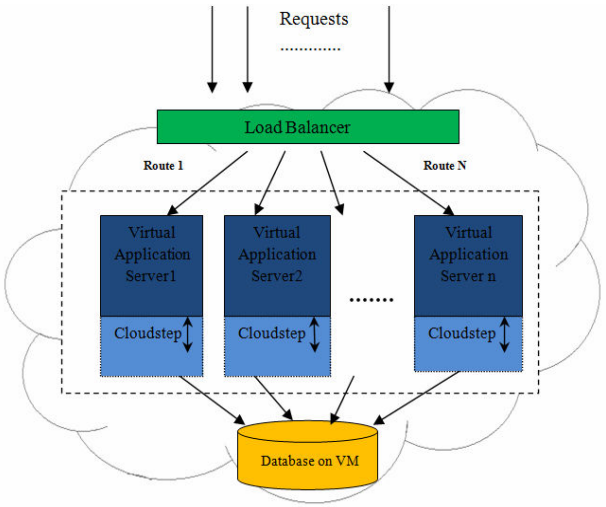

Fig. 3. Proposed Scaling Approach

In this cloud scaling environment firstly we go for scaling up of application using certain amount of cloudsteps. Resource consumption depends on the type and quantity of application. When the application starts extra unused cloudsteps are added or reduced automatically based on the increase or decrease of the load. Adding or reducing cloudsteps and also we can set the scalability limits for effective budget/cost control. If application starts requesting more and more resources, the running application server is completely loaded and might not be able to provide the required resources anymore so we opt new instance that is scale out. This scaling approach is one kind of diagonal scaling i.e. scale up and then out.

The work flow diagram of our proposed approach as shown in Fig 4. This is based on the scaling indicator CPU/RAM usages in each virtual machine instance in the Cloud. In this scenario initially capacities in Owned definition will be totally instantiated as VM instances and be started, then joined into the front end load balancer. If any instance required additional resources CPU or RAM automatically based on CPU or RAM threshold to handle load and if decreases remove additional resources from the back end application server. If application server is fully utilized and might not be able to provide the required resources anymore then a new application instance will be provisioned, started, and then added to the load-balancer. If there are instances with nil usages of resources and with at least one instance that has nil load, the idle instance will be removed from the load-balancer and be shutdown from the system. In each case, the load factors for all active instances will be recalculated and then applied to the load-balancer to re-distribute the request workloads to each instance evenly.

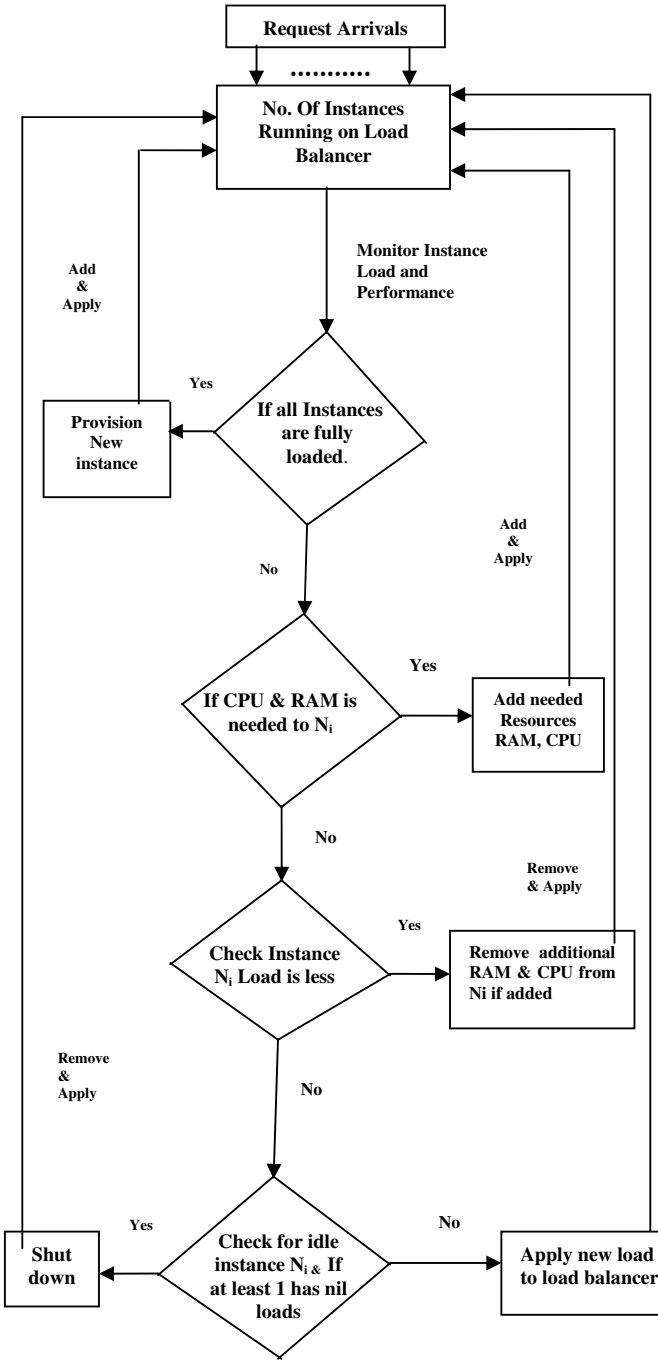

Fig. 4. Proposed Automatic Resource Provisioning Mechanism

V. CONCLUSION

Scalability is the best solution to increasing and maintaining application performance in cloud computing environments. But currently cloud scalability is restricted to scale out, and there is need of scaling up for handling most sudden, temporary peaks in application on cloud environment for reducing cost and higher resource utilization. In summary, we have designed a prototype of hybrid scaling scenario to address the dynamic scalability of cloud applications in which scalable applications is deployed at PaaS cloud for scale up and when application server fully utilized then go for scale out and future work is to validate the implemented features using, experiments.

REFERENCES

[1] G. Gruman, "What cloud computing really means", InfoWorld, Jan. 2009.
[2] R. Buyya, Y. S. Chee, and V. Srikumar, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities", Department of Computer Science and Software Engineering, University of Melbourne, Australia, pp. 9, July 2008.
[3] D. Chappell, "A Short Introduction to Cloud Platforms", David Chappell & Associates, August 2008.
[4] C. Braun, M. Kunze, J. Nimis, and S. Tai. Cloud Computing, *Web-based Dynamic IT-Services*. Springer Verlag, Berlin, Heidelberg, 2010.
[5] L. M. Vaquero, L. Rodero-Merino and Rajkumar Buyya, "Dynamically Scaling Applications in the Cloud," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, pp. 45–52, january 2011.
[6] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 50–55, 2009.
[7] L. Youseff, M. Butrico, and D. da Silva, "Toward a unified ontology of cloud computing," in *Proceedings of the Grid Computing Environments Workshop*, Austin, Texas, USA, pp. 1–10, November 2008.
[8] A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm, "What's inside the cloud? An architectural map of the cloud landscape," in *ICSE 2009: Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, Vancouver, Canada, pp. 23–31, May 2009.
[9] 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," *Future Gener. Comput. Syst.*, vol. 25, no. 6, pp. 599–616, 2009.
[10] Trieu C. chieu, Ajay Mohindra, Alexei A. Karve, Alla Segal, " Dynamic scaling of Web Applications in a Virtualized Cloud Computing Environment",IEEE International Conference on e-Business Engineering, pp. 281-286,2009.
[11] E. Knorr, "Software as a service: The next big thing", InfoWorld, March 2006.
[12] Comparing Amazon's and Google's Platform-as-a-Service (PaaS) Offerings, Enterprise Web 2.0, ZDNet.com.
[13] VMware Inc., "Understanding Full Virtualization, Paravirtualization, and Hardware Assist", VMware, 2007,
[14] M. Rosenblum, T. Garfinkel. —Virtual Machine Monitors: Current technology and future trends, Computer, 38(5), Los Alamitos, CA, IEEE Computer Society Press, pp. 39-47, 2005
[15] Hitesh A Bheda, Chirag S Thakur," Performance Optimization of Workload usage with Virtualization in Cloud Computing Environment", IRCTITCS, 2011.
[16] GoGrid cloud hosting, " Scaling your internet business", 2010.
[17] Ming Mao, Jie Li, Marty Humphrey, " Cloud Auto-scaling with Deadline and Budget Constraints",2011.
[8] Dominique Bellenger, Jens Bertram, Andy Budina, Benjamin Pfander," Scaling in Cloud Environments", Recent Researches in Computer Science,2011.
[19] Amazon.com. Auto Scaling, http://aws.amazon.com/autoscaling/
[20] Joyent white paper, " Performance and Scale in Cloud Computing", http://www.joyent.com/
[21] Autoscaling java application, http://www.jelastic.com/

Bhushan Lal Sahu received B.E. degree in Information Technology from CSVT University, Bhilai, Chhattisgarh, India in 2009. He is at present doing ME in Computer Technology and Application from Shri Shankaracharya College of Engineering and Technology, Bhilai affiliated to CSVTU, Bhilai, Chhattisgarh, India. His area of interest includes Cloud Computing and Computer networking.

Rajesh Tiwari is an Associate Professor of Computer Science and Engineering at Shri Shankaracharya College of Engineering and Technology, CSVT University, Bhilai, Chhattisgarh, India. He received the A.M.I.E. in Computer Science and Engineering and M.E degree in Computer Technology and Application from Shri Shankaracharya College of Engineering and Technology, Bhilai, Chhattisgarh, India in 2008. His research works including Parallel Computing and Cloud Computing.