Implementation of Template based Datacenter Broker Policy for Cloud to Improve the QoS

Seema Chowhan, Ajay Kumar, Shailiaja Shirwaikar

Abstract*: *Cloud computing is abstraction of web based services for businesses having dynamic requirements of resources. The cloud offers noteworthy benefits to business world by providing hardware set up, software and management of the system software. It puts in together elastic resources including hardware, software in virtual platform dynamically to meet computing need. The Quality of service parameters (QoS) comprises different parameters like, performance, availability and reliability, response time, throughput and bandwidth etc. There is need for proper monitoring, management of these parameters for effective cloud services and to maintain customer base. An efficient DCB policy reduce the overall execution time of the requested Cloudlets (Jobs/Tasks). An important policy of the Datacenter Broker (DCB) is binding of Cloudlets with an available VMs. For efficient load balancing there is need of proper allocation of cloudlets to the appropriate available VMs as per application requirement. In present study, we proposed a Template Based Resource Provisioning algorithm for effective resource utilization and allocation of Cloudlets to the available VMs in a Datacenter. The proposed method takes into consideration application requirement with its size (Cloudlet length) along with power and capacity of VMs. Experimental are performed using CloudSim under different workload scenario and results are obtained for comparison. The proposed algorithm creates performance supremacy over existing DCB algorithm for high workload.

Keywords: Quality of Service, Virtual Machines, Datacenter, Datacenter Broker.

I. INTRODUCTION

Computing technology has experienced a series of platform, software and environment changes over the past years. Cloud computing is a repository of heterogeneous resources for parallel and distributed computing to solve large scale problem over the Internet. Clouds can build with virtualized resources on the top of physicals resources over large data centers that are centralized or distributed. As there are millions of users using cloud based services with pay as you go manner, the load balancing is an important issue to be addressed by cloud provider. The load balancing at resource provisioning involves in management of huge pool of resources and providing these resources to customer as per their requirement. Cloud provider has the herculean task of satisfying the **elastic demand** of its users with an optimal investment in infrastructure. From cloud provider's perspective, load balancing is an important activity to achieve efficient utilization of resources. SLA based resource provisioning involves maintaining appropriate levels of various resources while meeting QoS attributes or minimization of SLA violations.

In present study, an improved Template Based Datacenter Broker algorithm is proposed for binding of a Cloudlet to a VM Template which comprised of small, medium and large VM with different configuration. VMs from VM template is allocated based on input load so that idle capacity is optimally utilized in a Datacenter as a result completion time of cloudlets are reduced to improve system utilization. CloudSim 3.0.3 simulation platform [5] is used for simulation experiments. DCB in CloudSim supports internal scheduling of cloudlets or tasks as well as for VM creation with First Come First Serve (FCFS) policy [3] and Round Robin (RR) scheduling strategies. There is need for improved scheduling strategy as FCFS and RR are time consuming as for bigger jobs it suffers from Long average waiting time. Proposed Template Based Datacenter Broker algorithm provides better result than the above-mentioned policies. It takes different workload scenarios as input which is product of cloudlet length and number of cloudlets to be processed so the lengthy (CPU intensive) cloudlets assigned to appropriate VMs as per given workload.

Section 2 describes the prior work done on load balancing and resource provisioning. Section 3 describes CloudSim and related terminology. In Section 4, presents proposed Template Based Resource Provisioning Method. Section 5 describes Comparison and Results. Section 6 conclude the research.

II. RELATED WORK

Cloud provider has to create an illusion of availability of unlimited computing resources to the end users on limited hardware and unpredicted request loads. The challenges for cloud computing provider is to allocate resources as per Service Level Agreements (SLAs) and performance of cloud system should be stable in any dynamic changes of workload as per SLA specified without effecting quality of service (QoS). By Byun et al [4] SLAs specify the resources and quality levels required for job execution so that it will minimize the cost from user perspective and to maximize the resource utilization from provider's perspective.

Manuscript published on January 30, 2020.
* Correspondence Author
Seema Chowhan *, Department of Computer Science, Baburaoji Gholap College Sangvi, Pune, India. Email: ssc_chow@yahoo.com
Ajay Kumar, Department of Computer Application, Jayawant Institute, Pune, India. Email: ajay19_61@rediffmail.com
Shailiaja Shirwaikar, Department of Computer Science, Savitribai Phule Pune University, Pune, India. Email: scshirwaikar@gmail.com

In such systems Quality of service parameters are availability, reliability, response time and throughput in contractual documents agreed between provider and customer called SLA [8], [9], [11], [16], and [18]. QoS parameters are very important in ranking service providers.

Wide variety of web applications in network are moving to the cloud platform for using its various services on demand basis in pay as you go manner. , which makes high performance load balancing important. Load balancing distributes workload traffic amongst multiple computing resources to balance the load. Load Balancing is gaining critical importance in a cloud computing environment. Efficient load balancing scheme guarantees provisioning of resources to cloud users on demand with efficient resource utilization. Load Balancing can be carried out both at resource provisioning level which is heavily dependent on the Service Level Agreement (SLA) and also at the resource utilization level. Load balancing at Resource provisioning involves the distribution of resources to different cloud users without increasing wasted capacities and yet maintaining required Quality of service. The consumer of the service may provide one or more Service level objectives depending on application specific requirements but the cloud provider has to translate them into low level technical attributes that can be monitored and controlled to achieve the higher level objectives. Efficient resource provisioning policies allow sharing of resources in Data Center to enhance the cloud performance. Resource provisioning that maintains quality of service with optimum resource utilization is one of the challenge. It is a multidimensional problem that can have issue based solution in the form of a set of services that helps to allocate and negotiate service level agreements (SLA) and design. Several research work have explored resource allocation using static and dynamic load balancing algorithms. Round robin [16], Min-Min and Max-Min [8] are commonly used static load balancing algorithms. Zhang [19] presented efficient load balancing mechanism using ant colony and complex network theory for under-load and over-load situations. Radojevic and Zagar [13] proposed Central load balancing decision model (CLBDM) with improvement on round robin algorithm which in turns makes use of session switching. Map Reduce-based load balancing technique for entity resolution was proposed by [10]. Nishant et al [12] recommended Ant Colony Optimization algorithm for distributing the workload among nodes at cloud for balancing the load.

Dynamic load balancing algorithms are more efficient with accurate result. In [1] an algorithm was proposed to get a suitable execution sequence for workflow activities and their recoverability by the adaptive scheduling algorithm (ASA) which considers resource allocation constraints and dynamic topology changes. The dynamic scheduling Earliest Deadline First (EDF) algorithm which is used in real time system scheduling for multiprocessor system for efficient load balancing [15]. Zhong et al [20] proposed resource scheduling strategy in cloud computing using genetic algorithm. In [14] make use of algorithms like round robin, equally spread current execution and Throttled load balancing algorithm to distribute the load across VM instance to check the performance parameter time and cost.

SLA is an important document for load balancing at resource provisioning level. The higher level attributes such as availability, reliability and low level attributes such as response time, throughput, latency time, downtime per week, Mean time to Repair (MTTR), Mean time between failures (MTBF) etc. are QoS parameters [17]. QoS of a cloud service varies drastically at small timescales, due to network traffic and load on the network. The proposed template based resource provisioning and utilization method and procedure overcomes the problem of over-provision and under-provision of resources at Data Center without effecting QoS and SLA on any workload. The method is driven by completion time as one of QoS parameter for different workload scenarios. Template-Based Resource Provisioning (TBRP) method was proposed] for resource provisioning and optimum utilization of idle capacity without breaking the SLA on variation of workload [6-7].

The need is to continuously measure,monitor and take majors to maximize SLA parameters.

III. CLOUDSIM AND RELATED TERMINOLOGY

University of Melbourne has developed CloudSim simulator for performing experiments in cloud computing environment. The simulator aims at simulating the construction of the infrastructure of cloud computing and comparing different service scheduling and allocation strategies [21].

Basic Terminology

• Cloudlets: Cloudlet is the task/job/process in CloudSim framework.

• Virtual Machines (VMs): It runs inside a Host, sharing host resources with other VMs. It processes cloudlets according to a resource allocation policy, defined by the Cloudlet Scheduler.

• Hosts: It is a physical entity in the cloud data center with huge bundle of resources.

• Data-centers: Data-center is collection of Cloud Resource with virtualized host-list.

• Data-center Brokers: Data-center Broker represents a broker performing on behalf of a cloud user.

• Processing Element (PE): Number of CPUs or cores assigned to the virtual machines. The processing/execution capability of the PE/CPU is represented using Million Instructions per Second (MIPs) unit.

• Bandwidth (Bw): It is a unit of communication capability of VMs.

In CloudSim environment, Data-center consists of fixed or different configuration of hosts or servers. The host in a data center is described with attributes like host-id, storage, RAM, processing power (MIPs), number of processing elements (PE) and bandwidth. The virtual machine is characterized by number of virtual machines (VM), VM Image size, VMRAM, VM bandwidth. The cloudlet which runs on VM are characterized by length (in terms of instructions), Input File Size and Output File Size. The mapping of task (cloudlet) to VM is possible with either time shared or space shared allocation policy. In space shared policy, required PEs are exclusively allocated to VM. If running VMs require more PEs than available, they have to wait in a queue until enough resources are free. In Time Shared policy available PEs are shared among running VMs and all the cloudlets run simultaneously.

Time shared policy is used in the following experiments. The software as a service requires parameters like MIPs, bandwidth and processing elements whereas IaaS requires RAM, storage and number of processing elements. We assume that cloud resource users only request for virtual nodes only, as we are working with SaaS during experiment. The experiments are performed on virtual nodes for multiple tasks which are in a same data center. Each task or application represents a user's request; which can dynamically increase as per the requirement. Effects of varying VM and cloudlet parameters are checked for task completion time. The CloudSim environment with data center, host, virtual machine and cloudlets are indicated in "Fig 1".

A host in a Datacenter that meets the memory, storage, and availability requirement is selected with default VM Allocation Policy for a VM deployment. In existing DCB policy Datacenter Broker allocates the cloudlets (CL) to the first available VM. As presented in "Fig2" there are six applications/ jobs and three VMs. According to existing policy, Cloudlet CL-1 binds to VM-1, CL-2 binds to VM-2, CL-3 binds to VM-3, CL-4 binds to VM-1, CL-5 binds to VM-2 and CL-6 binds to VM-3 in round robin manner.

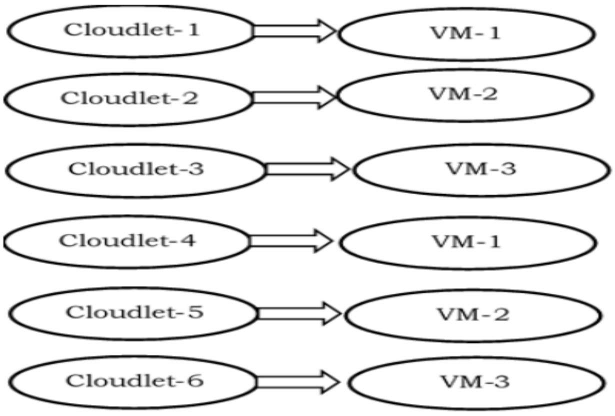

Fig.2. Cloudlet Binding with VMS

The existing Datacenter Broker Policy doesn't consider the processing capability of the VMs and type of workload before assigning Cloudlets to available VMs. Large size Cloudlets are frequently bind to the VMs with low MIPS rating (Low processing capacity VM) and therefore it suffers from longer completion time, waiting time and response time to execute the Cloudlets. Moreover there are chances of assigning least lengthy Cloudlets to most powerful VMs hence its resource utilization capacity gets wasted. So it reduces overall performance of the system.

IV. TEMPLATE BASED RESOURCE PROVISIONING (TBRP) ALGORITHM

In proposed Template Based resource provisioning algorithm modification is done in existing DCB policy in order to improve its performance [7].

Proposed TBRP method fulfills following objectives

1) Provisioning of resources as per workload requirements of user applications.

2) Avoid SLA violation as per completion time SLA clause designed.

3) Monitoring of resource utilization and minimization of resource utilization cost.

In this method resource provisioning and utilization strategy system is designed such that VMs are utilized sparingly at low workload and judiciously at high work load to maintain SLA by maintaining Quality of Service (QoS) levels. In TBRP method the entire resource are sliced into small, medium and large VM types and different combination of these VM types are designed as Templates. So that resource provisioning strategy periodically maps idle capacity into a set of VM templates for cost constraint and efficient resources utilization. The MIPs rating (processing capacity) of small-VM, medium-VM and large-VM in a VM template depends on capacity of Data-Center that will meet users SLA requirements and helps in managing overprovisioning and under provisioning. In case of customer requirement, data-center will provide available templates or custom made Templates to meet customer requirement. The method also utilizes full capacity of Data-Center to avoid over-provisioning and under-provisioning while giving attractive pricing for users. The Cloud service provider can also maximize profit without affecting the customer satisfaction. [19]

A. Workload (Cloudlet) Allocation Strategies

The Cloudlets are generated by users as per their workload requirement. The user workload that is product of Cloudlets and cloudlet length are submitted to the Datacenter Broker that has information regarding the processing capability of VMs. In proposed TBRP Algorithm workload are allocated to VMs correctly as per workload type and avoid SLA violation.

B. Working principle

In designing templates, apart from the configuration one need to define the SLA clauses in the QoS constraints. The parameters used in these can be derived by performing simulation experiments. Proposed Resource allocation policy by DCB is presented in Table-I.

Table- I: Resource allocation and utilization strategy [7]

Resource allocation and utilization strategy
Input: Predicted workload and the template configuration
QoS parameters in SLA/ VM template: MinCTvalue...LowWorkload...
Algorithm:
If predicted workload is less than LowWorkload
Start and use a single small VM
if completion time exceeds minCTvalue then
Completion Time SLA violation
Else
If predicted load is less than MedWorkload
Start and use both small and medium VM by
Distributing the load proportionately
If completion time exceeds medCTvalue then
Completion Time SLA violation
Else
If predicted load is less than MaxWorkload
Start and use all the three VMs by
Distributing the load proportionately
if completion time exceeds maxCTvalue then
Completion Time SLA violation
Else
Maxworkload SLA violation

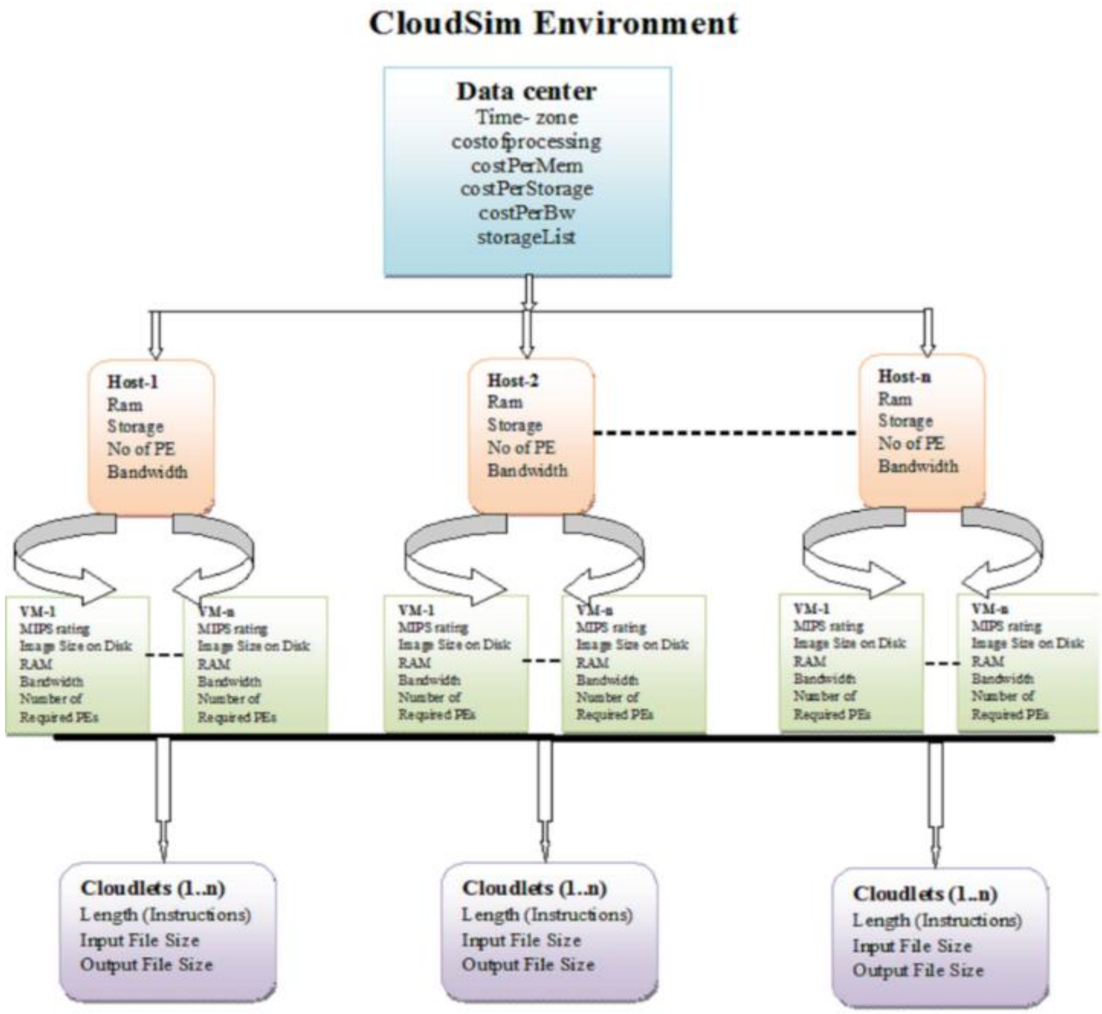

Fig.1. CloudSim Environment

The load is then proportionately divided among these VMs. The Resource utilization strategy thus uses provisioned VMs sparingly at the same time takes care that there are no SLA violations.

V. COMPARISON AND RESULTS

Consider a VM template of simple configuration comprising of one small, one medium and one large VM. The computing load depends on the application and varies at different point in time. In general four scenarios are considered. The application is having low workload that is when there are few tasks of small size. The workloads is medium when the both task size and number is same or one of them is small and the other is not very large.

The load reaches the pick when either large size tasks get executed or tasks increase in number. The load is extreme when both size of the task and number increase beyond limit. In the simulation experiment, workload for these four situations are executed on VM template of single small, medium, large VM with proposed TBDCB and also on VMs of same MIPs capacity as Template with existing DCB policy. In case of template the load is distributed proportionately.

Table-II: Comparison result of TBDCB and Existing DCB cloudlet allocation policies

Cloudlets (Number of task)	Cloudlet length (byte)	Workload (bytes)	Completion Time (Sec)	
			VM-Templates-SML	Existing Data-center Broker Policy
Low-workload				
2	100	200	2	0.83
4	100	400	3.99	1.28
6	100	600	5.96	1.43

8	100	800	7.95	2.11
10	100	1000	10	2.63
Medium-workload				
2	1000	2000	9.99	8.33
4	1000	4000	16.66	12.91
6	1000	6000	26.65	14.43
8	1000	8000	33.3	21.24
10	1000	10000	39.73	26.49
Large-workload				
2	10000	20000	66.66	83.33
4	10000	40000	99.96	129.16
6	10000	60000	149.99	144.43
8	10000	80000	199.99	212.49
10	10000	100000	249.99	264.98

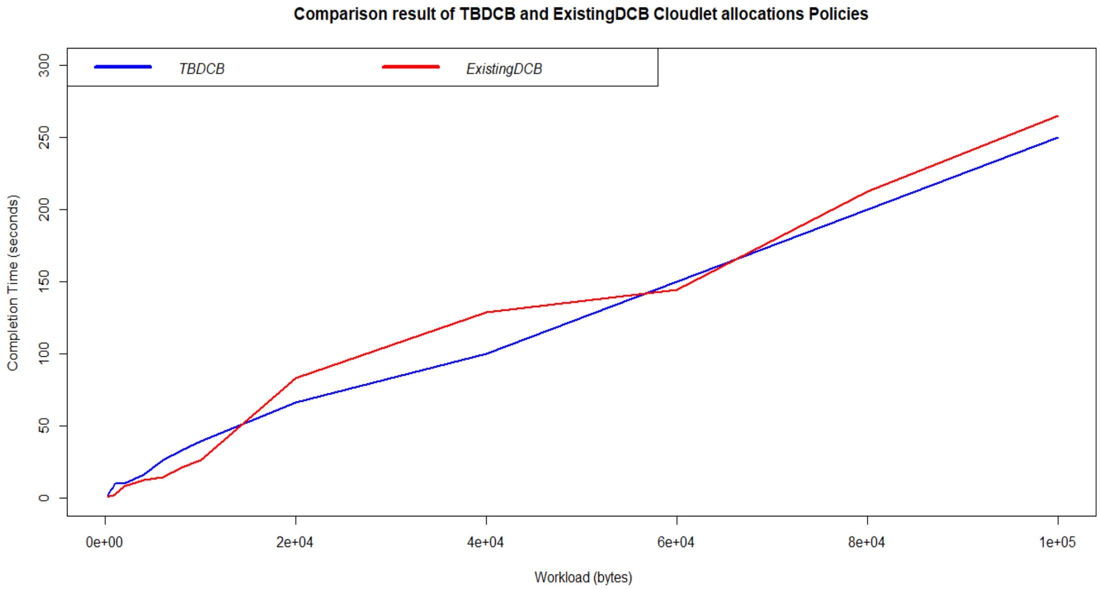

Fig.3. Comparison result of TBDCB and Existing DCB cloudlet allocation policies

Observation:

The comparison result of different workload scenarios are presented in Table-II and “Fig.3 “. With completion time of VMs of two different Cloudlet allocation policies it was observed that for low workload results are very close. For medium and large workload performance is better in proposed TBDCB policy as compared to existing DCB policy.

VI. CONCLUSION

The proposed Template Based Resource provisioning algorithm explores working of Data Center broker for allocation of workload as per user requirement to the different VMs inside the VM template. It distributing workload proportionately among VMs as per resource allocation and utilization strategy. It identifies the load intelligently with low, medium, high workload scenarios and start only small VMs for low workload, small and medium VMs for medium workload and all VMs at large workload. The proposed method provides better completion time and avoid SLA violation.

REFERENCES

1. Avanes, A. and Freytag, J.-C. 2008. Adaptive workflow scheduling under resource allocation constraints and network dynamics. Proceedings of the VLDB Endowment 1, 2, 1631–1637.
2. Bianco, P., Lewis, G. A., and Merson, P. 2008. Service level agreements in service-oriented architecture environments. Tech. rep., CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST.
3. Buyya, R., Ranjan, R., & Calheiros, R. N. (2009, June). Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities. In *2009 international conference on high performance computing & simulation* (pp. 1-11). IEEE.
4. Byun, E. K., Kee, Y. S., Kim, J. S., & Maeng, S. (2011). Cost optimized provisioning of elastic resources for application workflows. *Future Generation Computer Systems*, *27*(8), 1011-1026.

5. Calheiros, R.N., Ranjan, R., De Rose, C.A.F., Buyya, R.: CloudSim: A Novel Framework for modelling and Simulation of Cloud Computing Infrastructures and Services (2009) 19
6. Chowhan, S. S., Kumar. A, & Shirwaikar, S. (2019) .Measuring effectiveness of TBRP method for different capacity Data Centers. IPASJ International Journal of Information Technology (IIJIT), Vol.-7, Issue-5 6-16
7. Chowhan, S. S., Kumar. A, & Shirwaikar, S. (2019) .Template-Based Efficient Resource Provisioning and Utilization in Cloud Data-Center. International Journal of Computer Sciences and Engineering, Vol.-7, Issue-1 463-477.
8. Etminani, K. and Naghibzadeh, M. 2007. A min-min max-min selective algorihtm for grid task scheduling. In 2007 3rd IEEE/IFIP International Conference in Central Asia on Internet. IEEE, 1–7.
9. John, M., Gurpreet, S., Steven, W., Venticinque, S., Rak, M., David, H., Gerry, M., DI MARTINO, B., Le Roux, Y., John, M., et al. 2012. Practical guide to cloud service level agreements.
10. Kolb, L., Thor, A., and Rahm, E. 2012. Load balancing for map reduce based entity resolution. In 2012 IEEE 28th international conference on data engineering. IEEE, 618–629.
11. Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., and Leaf, D. 2011. Nist cloud computing reference architecture. NIST special publication 500, 2011, 1–28.
12. Nishant, K., Sharma, P., Krishna, V., Gupta, C., Singh, K. P., Rastogi, R., et al. 2012. Load balancing of nodes in cloud using ant colony optimization. In 2012 UKSim 14th International Conference on Computer Modelling and Simulation. IEEE, 3–8.
13. Radojevi´c, B. and ˇZagar, M. 2011. Analysis of issues with load balancing algorithms in hosted (cloud) environments. In MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, 416–420.
14. Sharma, T. and Banga, V. K. 2013. Efficient and enhanced algorithm in cloud computing. International Journal of Soft Computing and Engineering (IJSCE) ISSN, 2231–2307.
15. Singh, J. 2010. An algorithm to reduce the time complexity of earliest deadline first scheduling algorithm in real-time system. arXiv preprint arXiv:1101.0056. 9, 34
16. Sotomayor, B., Montero, R. S., Llorente, I. M., and Foster, I. 2009. Virtual infrastructure management in private and hybrid clouds. IEEE Internet computing 13, 5, 14–22.
17. Tang, B. and Tang, M. 2014. Bayesian model-based prediction of service level agreement violations for cloud services. In TASE. 170–176.
18. Wu, L., Garg, S. K., and Buyya, R. 2011. Sla-based resource allocation for software as a service provider (saas) in cloud computing environments. In Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. IEEE Computer Society, 195–204.
19. Zhang, Z. and Zhang, X. 2010. A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation. In 2010 The 2nd International Conference on Industrial Mechatronics and Automation. Vol. 2. IEEE, 240–243.
20. Zhong, H., Tao, K., and Zhang, X. 2010. An approach to optimized resource scheduling algorithm for open-source cloud systems. In 2010 Fifth Annual China Grid Conference. IEEE, 124–129.

AUTHORS PROFILE

Ms. Seema Chowhan is working as a faculty and head in subject of computer science in Baburaoji Gholap College Pune, India affiliated to Savitribai Phule Pune University, Pune. She has 18+ years of experience in teaching UG and PG courses. She has completed M.Phil (CS).Her research interests include Cloud Computing and Networking.

Dr. Ajay Kumar experience covers more than 26 years of teaching and 6 years of Industrial experience as IT Technical Director and Senior Software project manager. He has an outstanding academic career completed B.Sc. App. Sc. (Electrical) in 1988, M.Sc. App.Sc. (Computer Science-Engineering and Technology) in1992 and PhD in1995. Presently, working as Director at JSPMs Jayawant Technical Campus, Pune (Affiliated to Pune University). His research areas are Computer Networks, Wireless and Mobile Computing, Cloud computing, Information and Network Security. There are 74 publications at National and International Journals and Conferences and also worked as expert, appointed by C-DAC to find Patent-ability of Patent Applications in ICT area. Six commercial projects are completed by him for various companies/ Institutions. He holds variety of imperative position like Examiner, Member of Board of Studies for Computer and IT, Expert at UGC.

Dr. Shailaja Shirwaikar has a PhD in Mathematics of Mumbai University, India and worked as Associate Professor at Department of Computer Science, Nowrosjee Wadia College affiliated to Savitribai Phule Pune University, Pune for last 27 years. Her research interests include Soft Computing, Big Data Analytics, Software Engineering, machine learning and Cloud Computing.