Load Balancing Approach in Cloud Computing

Majid Mehmood[1], Kinza Sattar[2], Asif Hussain Khan[3] and Mujahid Afzal[3]

[1]Department of Information Technology, University of Lahore, Gujrat, Pakistan
[2]Department of Computer Science, COMSATS, Abbottabad, Pakistan
[3]Department of Information Technology, University of Lahore, Gujrat, Pakistan

Abstract

Cloud computing is a utility to deliver services and resources to the users through high speed internet. It has a number of types and hybrid cloud is one of them. Delivering services in a hybrid cloud is an uphill task. One of the challenges associated with this paradigm is the even distribution among the resources of a hybrid cloud, often refereed as load balancing. Through load balancing resource utilization and job response time can be improved. This will lead to better performance results. Energy consumption and carbon emission can also be reduced if the load is evenly distributed. Hence, in this paper we have conducted a survey of the load balancing algorithms in order to compare the pros and cons of the most widely used load balancing algorithms.

Keywords: Cloud; Cloud computing; Load balancing

Introduction

Cloud computing

Cloud computing is a utility to deliver services and resources to the users through high speed internet [1]. It has gained immense popularity in recent years. These cloud computing services can be used at individual or corporate level [2,3]. Cloud computing can be summarized as a model that gives access to a pool of recourses with minimal management effort [4].

Types of clouds

Clouds can be classified as private, public and hybrid [6,7,8] on the basis of their architecture. It provides three types of services 1. Infrastructure as a Service (IAAS), that provides the infrastructure a user demands like routers. 2. Software as a Service (SAAS), delivers software services like Google Apps. 3. Platform as a Service, PAAS, as the name suggests provides platforms for program development for example Google's App Engine [5].

Private cloud: A cloud used only within an enterprise is referred as a private cloud [6]. It can also be addressed as internal cloud [8]. They are managed by the organization itself.

Public cloud: A cloud that is made available to the users around the globe through an Internet access is called a public cloud [6]. Organizations providing such cloud services include Google Docs [9], Microsoft's Windows Azure Platform [10], Amazon's Elastic Compute Cloud and Simple Storage Services [11], IBM's Smart Business Services [12].

Hybrid cloud: A union of private and public clouds forms another type of cloud referred as hybrid cloud. As one part of it is private, it is considered to be more secure but designing a hybrid cloud is a challenging job because of the complexities involved in the design phase [8]. The major issues linked with them are that of interoperability and standardization [13]. They are costly as compared to the aforementioned types but it has their best features combined [14].

Benefits of hybrid clouds

Hybrid cloud model provides a seamless integration of public and private infrastructure which allows the use of public resources when local resources run out. The term normally used to refer to this state is called "cloud bursting". An elastic environment is created this way. Some benefits of hybrid clouds are listed as follows [15,16].

Optimal resource utilization: One strategy for handling peak load is to extend the infrastructure that could handle the load when it crosses the average load. The peak load is almost ten times the average load. That's why this strategy would be costly. In hybrid cloud environment, the public cloud is used at the time of peak loads. That saves from unnecessary investment on infrastructure expansion.

Risk transfer: It also reduces the risk of downtime. If some problem occurs in the private cloud the load is transferred to the public infrastructure. This way, hybrid clouds help in increasing the uptime.

Availability: Hybrid clouds provide availability without having to provide redundant and geographically distributed infrastructure.

Reduction in hardware cost: With the help hybrid clouds the need to expand the data centers has reduced as the peak load is faced only occasionally, adding infrastructure to cater this requirement would be a waste of money. As mention before, peak load is ten times the average load.

Better QoS: By using proper scheduling techniques, hybrid clouds can help meet the desired deadlines and execution times hence QoS and SLAs can be delivered as promised.

Challenges and issues of hybrid clouds

However, many challenges are also associated with hybrid clouds as elaborated in [15-17]. Some are briefly discussed here:

Interoperability and portability: Vendor lock-ins is always given importance whenever outsourcing is discussed. Interoperability and Portability is an open research problem in hybrid clouds for the researchers. Interoperability defines the way how communication between different clouds would occur. If we take the example of Amazon and Google, the image of Windows used is the same on both, no change is required, this is called interoperability. On the other hand, portability refers to the capability of a cloud, to move data and

application from one cloud to another. This becomes a possibility only if the dependencies are removed.

Cost: In these environments, on one hand, private infrastructure is to be managed, while on the other hand, you are charged on the basis of pay-per-use for using the public resources. This makes predicting the overall cost an uphill task.

Security: For using public cloud resources certain SLAs are settled first and a lot of trust is placed in the public cloud. Additional security measures are to be taken along with the company's firewalls. That is why security is one of the primary concerns in hybrid cloud environment. To ensure secure computation, some security issues are given prime importance. The list includes: Identification and authentication, authorization, and confidentiality etc. [18].

Reliability: As the communication between a private and public cloud occurs through a network connection, the availability of connection is again an issue as connection often breaks. Are these connections secure or not and would the migration of tasks to the public cloud actually help reduce the response time or not, are the questions that need to be addressed. So ensuring reliability is another challenge.

Monitoring: Organizations monitor the cloud services to ensure the performance is not compromised in any situation. In hybrid clouds, along with monitoring the private cloud, public clouds also need to be monitored.

Denial of service: Another challenge that is inspected by the researchers is the denial of service (DoS) in cloud computing environments. As in normal clouds and even in the hybrid environments resources are allocated dynamically, how would these clouds respond to a DoS attack, is a question given a lot of importance in the recent years. In hybrid clouds if resources are not available to the executing tasks, those tasks are forwarded to the public clouds but in this case the strategy discussed won't be a feasible solution. Finding a solution to this problem is a burning challenge for the researchers.

Load balancing: Load balancing is also one of the main challenges faced in hybrid cloud computing, as there is a need for an even and dynamic distribution of load between the nodes in private and public clouds.

In distributed systems load balancing is defined as the process of distributing load among various nodes to improve the overall resource utilization and job response time. While doing so, it is made sure that nodes are not loaded heavily, left idle or assigned tasks lesser then its capacity [19]. It is ensured that all the nodes should be assigned almost the same amount of load [20].

If resources would be utilized optimally, performance of the system will automatically increase. Not only this, the energy consumption and carbon emission will also reduce tremendously. It also reduces the possibility of bottleneck which occurs due to the load imbalance. Furthermore, it facilitates efficient and fair distribution of resources and helps in the greening of these environments [21,22].

Load balancing algorithms are classified into categories for the ease of understanding. That helps in identifying a suitable algorithm in the time of need. A detailed view of classification is presented below [23].

Related Work

With the emergence of hybrid clouds, the idea of balancing the load between the public and private clouds has gained immense popularity. That is why a lot of research in now being conducted to facilitate this

concept. Load balancing has always been given prime importance in cloud environment. Lately, researchers have started expanding this idea to the hybrid clouds as well in order to balance load at peak times while meeting the promised QoS and SLAs.

Load balancing strategies for clouds

Load balancing algorithms can be broadly categorized into static and dynamic load balancing algorithms.

Static load balancing algorithms: Gulati et al. [24] claimed that in cloud environment a lot of work is done on load balancing in homogeneous resources. Research on load balancing in heterogeneous environment is given also under spot light. They studied the effect of round robin technique with dynamic approach by varying host bandwidth, cloudlet long length, VM image size and VM bandwidth. Load is optimized by varying these parameters. CloudSim is used for this implementation.

Dynamic load balancing algorithms: A hybrid load balancing policy was presented by Shu-Ching et al. [25]. This policy comprises of two stages 1) Static load balancing stage 2) Dynamic load balancing stage. It selects suitable node set in the static load balancing stage and keeps a balance of tasks and resources in dynamic load balancing stage. When a request arrives a dispatcher sends out an agent that gathers nodes information like remaining CPU capacity and memory. Hence the duty of the dispatcher is not only to monitor and select effective nodes but also to assign tasks to the nodes accordingly. Their results showed that this policy can provide better results in comparison with min-min and minimum completion time (MCT), in terms of overall performance.

Another algorithm for load balancing in cloud environment is ant colony optimization (ACO) [26]. This work basically proposed a modified version of ACO. Ants move in forward and backward directions in order to keep track of overloaded and under loaded nodes. While doing so ants update the pheromone, which keeps the nodes' resource information. The two types of pheromone updates are 1) Foraging pheromone, which is looked up when an under loaded node is encountered in order to look for the path to an over loaded node. 2) Trailing pheromone is used to find path towards an under loaded node when an over loaded node is encountered. In the previous algorithm ants maintained their own result sets and were combined at a later stage but in this version these result sets are continuously updated. This modification helps this algorithm perform better.

Genetic algorithm [27] is also a nature inspired algorithm. It is modified by Pop et al. [28], to make it a reputation guided algorithm. They evaluated their solution by taking load-balancing as a way to calculate the optimization offered to providers and makespan as a performance metric for the user.

Another such algorithm is the bees life algorithm (BLA) [29], which is inspired by bee's food searching and reproduction. This concept is further extended to specifically address the issue of load balancing in [30]. The Honey bee behavior inspired load balancing (HBB-LB) algorithm basically manages the load across different virtual machines for increasing throughput. Tasks are prioritized so that the waiting time is reduced when they are aligned in queues. The honey bee foraging behavior and some of its variants are listed in [31].

Comparison

A comparative study of different load balancing algorithms is presented in [32]. Load balancing is not only required for meeting

users' satisfaction but it also helps in proper utilization of the resources available. The metrics that are used for evaluating different load balancing technologies are: throughput, overhead associated, fault tolerance, migration time, response time, resource utilization, scalability, and performance. According to this study, in honeybee foraging algorithm, throughput does not increase with the increase in system size. Biased random sampling and active clustering do not work well as the system diversity increases. OLB + LBMM shows better results than the algorithms listed so far, in terms of efficient resource utilization. The algorithm Join-Idle-Queue can show optimal performance when hosted for web services but there are some scalability and reliability issues that make its use difficult in today's dynamic-content web services. They further added that min-min algorithm can lead to starvation. They concluded that one can pick any algorithm according to ones needs. There is still room for improvement in all of these algorithms to make them work more efficiently in heterogeneous environments while keeping the cost to a minimum. A somewhat similar analysis of load balancing algorithms is presented by Daryapurkar et al. [33] and Rajguru and Apte [34] as well. Different scheduling algorithms for the hybrid clouds compared by Bittencourt et al. [35], highlights that the maxspan of these algorithms widely depend on the bandwidth provided between the private and public clouds. The channels are usually part of the internet backbone and their bandwidth fluctuates immensely. This makes the designing of the communication aware algorithms quite challenging.

Load Balancing Strategies in Hybrid Clouds

Zhang et al. [36] proposed a design for hybrid cloud is. It allows intelligent workload factoring by dividing it into base and trespassing load. When a system goes into a panic mode the excess load is passed on to the trespassing zone. Fast frequent data item detection algorithm is used for this purpose. It makes use of the least connections balancing algorithm and the Round-Robin balancing algorithm as well. Their results show that there is a decrease in annual bills when hybrid clouds are used. Buyya et al. [37] proposed a concept of federated cloud environment, to maintain the promised QoS even when the load shows a sudden variation. It supports dynamic allocation of VMs, Database, Services and Storage. That allows an application to run on clouds from different vendors. In Social Networks like Facebook, load varies significantly from time to time. For such systems this facility can help scale the load dynamically. No cloud infrastructure provider can

have data centers all around the globe. That's why to meet QoS, any cloud application service provider has to make use of multiple cloud providers. For implementation purpose they used Cloud Sim Tool kit. They made a comparison between federated and non federated cloud environments. Their results showed a considerable gain in performance in terms of response time and cost in case of the former. The turnaround time is reduced by 50% and the make span improves by 20%. Although the overall cost increases with the increase in the public cloud utilization but one has to consider that such peak loads are faced occasionally which makes it acceptable.

Task scheduling plays a vital role in solving the optimization problem in hybrid clouds. A graph-based task scheduling algorithm is proposed by Jiang et al. for this purpose [38]. In order to reduce the cost to a minimum value, like other algorithms, it makes use of the public resources along with the private infrastructure. The key stages of this algorithm are 1) Resource discovery and filtering, for the collection of the status information of the resources that are discovered. 2) Resource selection, this algorithm's main focus is on this stage as this is the decision making stage. Resources are picked keeping in view the demand of the tasks to be performed. 3) Task submission, once the resources are selected the tasks are assigned accordingly. A bipartite graph G=(U,V,E) is used to help elaborate this concept, where U is used for private or public Virtual Machines, V is for the tasks, and E denotes the edges in between. Cloud Report and Cloud Sim 3.0 are used for evaluating this algorithm. Their results showed a 30 % decrease in cost as compared to a non hybrid environment. For improving these figures even more, disk storage and network bandwidth need to be considered as well.

Another algorithm, adaptive-scheduling-with-QoS-satisfaction (AsQ) [39], for the hybrid cloud is proposed that basically reduces the response time and helps increase the resource utilization. To fulfill this goal several fast scheduling strategies and run time estimations are used and resources are then allocated accordingly. If resources are used optimally in the private clouds, the need for transferring tasks to the public clouds decreases and deadlines are fulfilled efficiently but if a task is transferred to the public cloud, minimum cost strategy is used so that the cost of using a public cloud can be reduced. The size of the workload is specially considered in this regard. Their results show that As Q performs better compared to the recent algorithms of similar
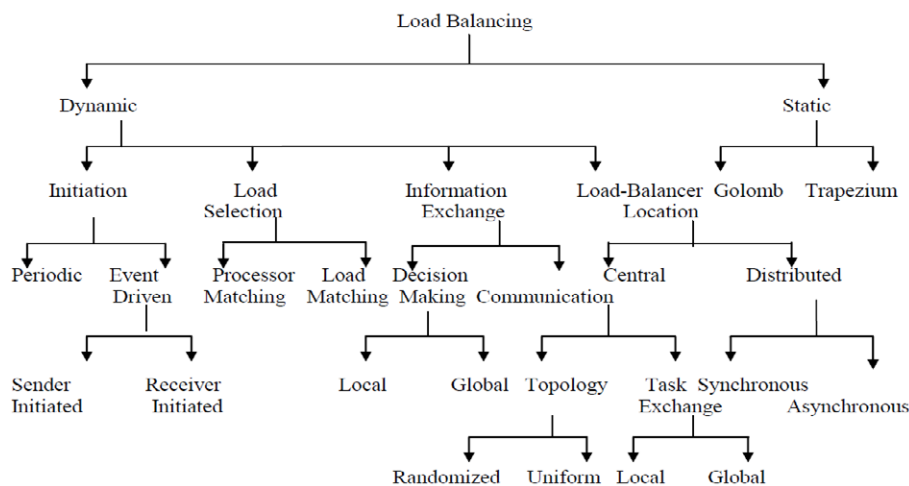

Figure 1: Classification of load balancing algorithms.

Metrics	Honeybee Scheduling	Biased random Sampling	Active Clustering	OLB+ LBMM	Join Idle Queue	Min-Min	Min-Max
Throughput	✗	✗	✗	✗	✗	✓	✓
Overhead	✗	✓	✓	✗	✓	✓	✓
Fault Tolerance	✗	✗	✗	✗	✗	✗	✗
Migration Time	✗	✗	✓	✗	✗	✗	✗
Response Time	✗	✗	✗	✗	✓	✓	✓
Resource Utilization	✓	✓	✓	✓	✗	✓	✓
Scalability	✗	✗	✗	✗	✗	✗	✗
Performance	✗	✓	✗	✓	✓	✓	✓

Figure 2: Comparison of the existing load balancing techniques.

nature in terms of task waiting, execution and finish time. Hence it provides better QoS.

Picking the best resources from the public cloud is a serious concern in hybrid clouds. The Hybrid Cloud Optimized Cost (HCOC) [40], is one such scheduling algorithm. It helps in executing a workflow within the desired execution time. Their results have shown that it reduces the cost while meeting the desired goals. Gives better results in comparison with the other greedy approaches. There is another approach [41], which also deals with directed acyclic graphs (DAG) as in study by Bittencourt and Madeira [40]. It uses integer linear program (ILP) for the workflow scheduling n SaaS/PaaS clouds with two levels of SLA, one with the customer one for the provider. This work can be extended by considering multiple workflows and fault tolerance in view.

Gupta et al. [42], contributed that there are a number of load balancing algorithms that basically help in avoiding situations where a single node is loaded heavily and the rest are either idle or have lesser number of tasks when in reality they can afford to deal with a lot more. But what is overlooked in most of these algorithms is the trust and reliability of the datacenter. A suitable trust model and a load balancing algorithm are proposed. They used VMMs (Virtual Machines Monitors) to generate trust values on the basis of these values nodes are selected and the load is balanced.

A virtual infrastructure management tool is offered by Hoecke et al. [43], that helps to set-up and manage hybrid clouds in an efficient way. This tool automatically balances load between the private and public clouds. It works at the virtual machine level. This tool has two parts 1) a proxy, where different load balancing algorithms are implemented like weighted round robin and forwarding requests to appropriate VMs, and on the other hand a management interface is designed that visualizes the hybrid environment and manages it too for example it can start and stop VMs, can form clusters of VMs, and can also manage the proxy remotely. It can be improved further by using a more efficient algorithm on the proxy for balancing load in a more convenient way.

In workflow applications [44], the cost of execution is kept to a minimum level by allocating the workflow to a private cloud but in case of peak loads, resources from the Public cloud need to be considered as well. As meeting the deadlines is a primary concern in workflow applications. By using cost optimization, this algorithm decides which

resources should be leased from the public cloud for executing the task within the deadline. In this algorithm workflow is divided into levels and scheduling is performed on each level. It uses the concept of sub-deadlines as well. That helps in finding the best resources in public cloud in terms of cost while keeping in view that the workflows are executed within the deadlines. Although the make span of level based approach is 1.55 times higher than the non level based approach, its cost is three times lower. In comparison with min-min, its make span is double but it costs three times lesser. This makes the proposed level based approach better as it costs less and meets the deadlines too although its make span is higher but it finishes the assigned tasks within the deadline.

Conclusion

Cloud computing is a utility to deliver services and resources to the users through high speed internet. It has a number of types and hybrid cloud is one of them. As one part of it is private, it is considered to be more secure but designing a hybrid cloud is a challenging job because of the complexities involved. Some benefits of hybrid clouds are optimal resource utilization, risk transfer, availability, reduction in hardware cost and better QoS. However, many challenges are also associated with hybrid clouds as elaborated. Some of them are interoperability and portability, cost, security, reliability, monitoring, denial of service, load balancing.

Load balancing algorithms can be broadly categorized into static and dynamic load balancing algorithms. A comparative study of different load balancing algorithms is presented. Load balancing is not only required for meeting users' satisfaction but it also helps in proper utilization of the resources available. The metrics that are used for evaluating different load balancing technologies are: throughput, overhead associated, fault tolerance, migration time, response time, resource utilization, scalability, and performance. According to this study, in honeybee foraging algorithm, throughput does not increase with the increase in system size. Different load balancing strategies in hybrid cloud are also discussed. A concept of federated cloud environment is proposed to maintain the promised QoS even when the load shows a sudden variation. Task scheduling plays a vital role in solving the optimization Problem in hybrid clouds. Another algorithm, adaptive-scheduling-with-QoS-satisfaction (AsQ) for the

hybrid cloud is proposed that basically reduces the response time and helps increase the resource utilization.

References

1. Cisco (2009) Cisco visual networking index: Forecast and methodology, 2009-2014.White paper.

2. Weiss A (2007) Computing in the clouds. Networker 11: 16-25.

3. Hayes B (2008) Cloud computing. Commun ACM 51: 9-11.

4. Mell P, Grance T (2009) Draft NIST Working Definition of Cloud Computing. Nat Inst Standards Technol.

5. Tao Z, Long J (2011) The Research and Application of Network Teaching Platform Based on Cloud Computing. Inter J Info Edu Tech 1.

6. Aditya S (2015) Top 5 Best Web Hosting Service Providers in 2015. Open Cloud Manifesto.

7. Armbrust M, Fox A, Griffith R, Joseph A D, Katz R H, A et al. (2009) Above the clouds: A Berkeleyview of cloud computing. Electr Eng Comput Sci Dept Univ California, Berkeley CA Tech.

8. Zhang Q, Cheng L, Boutaba R (2010) Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications 1: 7-18.

9. https://www.google.co.in/docs/

10. Azure Services Platform.

11. Amazon Web Services.

12. IBM Smart Business Services.

13. Dillon T, Wu C, Chang E (2010) Cloud Computing: Issues and Challenges. International Conference on Advance Information Networking and Applications IEEE.

14. Adler B (2012) Designing Private and Hybrid Clouds. Architectural Best Practices.

15. M Armbrust, A Fox, R Griffith, AD Joseph, R Katz, et al. (2009) Above the Clouds: A Berkeley View of Cloud Computing. Technical Report.

16. Heckel PC (2010) Hybrid Clouds: Comparing Cloud Toolkits. Seminar Paper Business Informatics, University of Mannheim.

17. Aubert BA, Patry M, Rivard S (2005) A Framework for Information Technology Outsourcing Risk Management. Database for Advances In Information Systems 36: 9- 28.

18. Tripathi A, Jalil MS (2013) Data Access and Integrity with authentication in Hybrid Cloud. Oriental International Journal of Innovative Engineering Research.

19. Rimal, Prasad B, Choi E, Lumb V (2009) A taxonomy and surveyof cloud computing systems. Proceedings of 5th International Joint Conference on INC, IMS and IDC, IEEE .

20. Sinha PK (1997) Distributed operating Systems Concepts andDesign. IEEE Computer Society Press.

21. Toledo RM, Gupta P (2010) Green data center: howgreen can we perform. Journal of Technology Research, Academic and Business Research Institute 2: 1-8.

22. Alakeel AM (2010) A Guide to dynamic Load balancing in Distributed Computer Systems. International Journal of Computer Science and Network Security 10: 153-160.

23. Khan Z, Singh R, Alam J, Saxena S (2011) Classification of Load Balancing Conditions for parallel and distributed systems. International Journal of Computer Science Issue 8: 411- 419.

24. Gulati A, Chopra RK (2013) Dynamic Round Robin for Load Balancing in a Cloud Computing, International Journal of Computer Science and Mobile Computing 2: 274-278.

25. Wang SC, Chen CW, Yan KQ, Wang SS (2013) The Anatomy Study of Load Balancing in Cloud Computing Environment. The Eighth International Conference on Internet and Web Applications and Services 230-235.

26. Nishant K, Sharma P, Krishna V, Gupta C, Singh KP, et al. (2012) Load Balancing of Nodes in Cloud Using Ant Colony Optimization. Computer Modelling and Simulation 3-8.

27. Whitley D (1994) A generic algorithm tutorial. Statistics and Computing 4: 65-68.

28. Pop F, Cristea V, Bessis N, Sotiriadis S (2013) Reputation guided Genetic Scheduling Algorithm for Independent Tasks in Inter-Clouds Environments. Advanced Information Networking and Applications Workshops.

29. Bitam S (2012) Bees Life Algorithm for Job Scheduling in Cloud Computing. International Conference on Communications and Information Technology 186-191.

30. Babu LD D, Krishna VP (2013) Honey bee behavior inspired load balancing of tasks in cloud computing environments. Applied Soft Computing 13: 2292-2303.

31. Suresh M, Ullah Z S, Kumar S (2013) An Analysis of Load Balancing in Cloud Computing. International Journal of Engineering Research and Technology 2: 428-433.

32. Sran N, Kaur N (2013) Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing. International Journal of Engineering Science Invention 2: 60-63.

33. Daryapurkar A, Deshmukh MV (2013) Efficient Load Balancing Algorithm in Cloud Environment. International Journal of Computer Science and Applications 6: 308-312.

34. Rajguru AA, Apte SS (2012) A Comparative Performance Analysis of Load Balancing Algorithms in Distributed System using Qualitative Parameters. International Journal of Engineering Research and Technology 1: 175-179.

35. Bittencourt LF, Madeira ERM, Fonseca N (2012) Scheduling in hybrid clouds. Communications Magazine, IEEE.

36. Zhang H, Jiang G, Yoshihira K, Chen H, Saxena A (2009) Intelligent Workload Factoring for A Hybrid Cloud Computing Model. Services World Conference.

37. Buyya R, Ranjan R, Calheiros RN (2010) Inter Cloud: Utility-Oriented Federation of Cloud Computing Environments for Scaling of Application Services. Algorithms and Architectures for Parallel Processing 1: 13-31.

38. Jiang WZ, sheng ZQ (2012) A New Task Scheduling Algorithm in Hybrid Cloud Environment. Cloud and Service Computing IEEE 45-49.

39. Wang WJ, Chang YS, Lo WT, Lee YK (2013) Adaptive scheduling for parallel tasks with QoS satisfaction for hybrid cloud environments. The journal of Super-Computing 66: 783-811.

40. Bittencourt LF, Madeira ERM (2011) HCOC: A Cost Optimization Algorithm for Workflow Scheduling in Hybrid Clouds. Journal of Internet Services and Applications, Springer 2: 207-227.

41. Genez TAL, Bittencourt LF, Madeira ERM (2012) Workflow Scheduling for SaaS /PaaS Cloud Providers Considering Two SLA Levels. Network Operations and Management Symposium 906-912.

42. Gupta P, Goyal MK, Kumar P (2013) Trust and Reliability based Load Balancing Algorithm for Cloud IaaS. Advance Computing Conference IEEE.

43. Hoecke SV, Waterbley T, Deneut JDT, Gelas JD (2011) Efficient Management of Hybrid Clouds. The Second International Conference on Cloud Computing, GRIDs, and Virtualization 167-172.

44. Chopra N, Singh S (2013) Deadline and Cost based Workflow Scheduling in Hybrid Cloud, International Conference on Advances in Computing. Communications and Informatics, IEEE.