INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING & TECHNOLOGY
VOLUME 25 ISSUE 1 – JULY 2017 - ISSN: 2349 – 9303

# Issues and Challenges in Minimizing the Response Time in Cloud Service

**G. Justy Mirobi**[1]
*Department Of Computer Science*
*Bharathiar University*
*georgejustymirobi1@gmail.com*

**L.Arockiam**[2]
*Department Of Computer Science*
*St. Joseph's College(Autonomous)*
*larockiam@yahoo.co.in*

**Abstract**— The cloud services are "Pay – Per – Use" policy over the internet. In cloud computing, the scheduling mechanism plays an important role for providing the services based on higher priority of the request. The scheduling mechanism is necessary for minimizing the response time, immediate switching time, complete utilization of resources, high bandwidth and for better performance of the service. This paper presents the challenges and requirements, subscription based cloud services, issues, and challenges. For the complete utilization of resources in cloud environment, the providers are allowing the oversubscription of resources. Therefore, the cloud resources are utilized in an efficient manner without any idle or waste of resources. Because of oversubscription, the workload is increased and there will be congestion in providing the service, therefore the response time will be maximized. To provide the best quality of service (QOS), the scheduling and workload should be analyzed from a capacity planning and resource provisioning perspective, for minimizing the response time.
**Index Terms**— Congestion, overload, oversubscription, scheduling, virtualization technique.

## 1.INTRODUCTION

Scheduling the jobs is used for rectifying the priority based issues in providing the resources. The degree of multiprogramming is controlled by scheduling the tasks. Scheduling the jobs is also one of the reasons for the maximum level of the response time. The scheduling mechanism is the solution to avoid deadline based jobs. The main function of the scheduling process is to select the task from the submitted request and send it to the system to take the decision for handling the selected task. Next, the cloud services are provided to the customers by using virtualization technique; it is showing that the unlimited amount of resources provided to each customer using virtualization. A single physical machine is divided based on the different type of resources and a set of virtual machines are created for the different type of resources. Sometimes, in the virtualization technique, a high amount of resources will be idle; the oversubscription will be the solution to avoid the idle level, waste of resources and for the standard cloud usage of resources in an efficient manner. Therefore the cloud resources are utilized in an efficient manner without any idle or waste of resources. Moreover, the amount of profit will be increased. The goal of oversubscription is the full utilization of resources and increasing the profit level. But the overload might occur in cloud environment due to oversubscription of the resources. The cloud resources might include disk, memory, CPU and other resources. The oversubscription is also one of the reasons for the maximum level of the response time. Next, in cloud environment, congestion is the situation of excessive resource allocation in providing the resources to the customers. The providers are providing the resources and expand the feature of oversubscription of the cloud resources for avoiding the waste, increasing the profit level and the complete utilization of resources. In congestion, the allocation of resources surpasses the actual capacity of the cloud resources, the load is increased and the bottle neck problem occurs. When the load is higher than the capacity, then the congestion problem occurs. This is also one of the reasons for the maximum level of the response time.

## 2.SUBSCRIPTION BASED CLOUD SERVICES

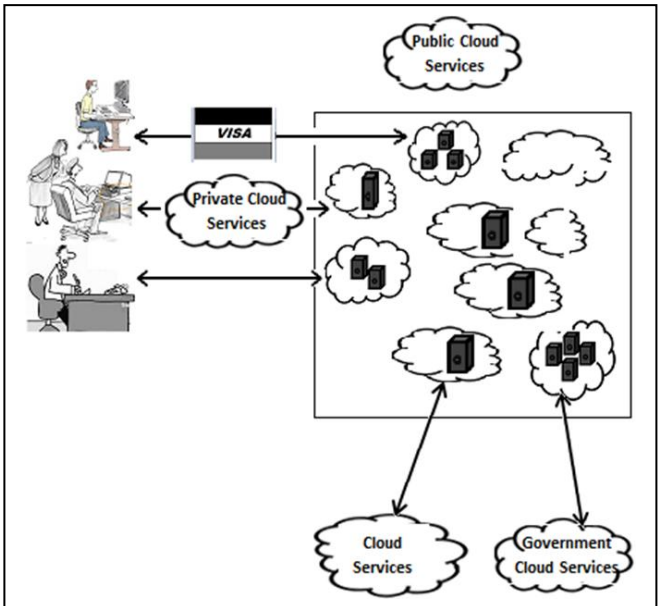

Fig. 1. Subscription based cloud services, in cloud computing.

Fig. 1. Shows the subscription based cloud services, in cloud computing, the users requests are defined in the Service Level Agreement SLA and the user expects that the required service to be provided by the service providers as defined in the SLA; the user pay the amount for the service which is provided. The infrastructure and services of cloud computing will be available on cloud servers, which can be provided to anywhere in the world, so that wherever the companies or individuals require the service can access the service anywhere in the world. In the cloud computing, the new approach for the usage of resources that are computational power, SAAS, PAAS and IAAS plays a vital role. The cloud computing provides the technology for establishing the connection dynamically to the data centers by combining the services of networked – virtual machines.

To keep the backup copy of the resources, make sure the reliability in case of single physical server, failure, the service providers are construction the distributed data centers at various locations in the world. So that workload is allotted to various data centers in the worldwide and save the response time by quick response. The servers in the data centers provide high availability, high reliability, high throughput, and high performance than the traditional computing services. Over subscription leads the problem of overload and unavailability in providing the resources from the cloud server. The SLA violations might be occurred and the response times will a maximum level because of improper scheduling, over subscription and congestion, therefore it is difficult to satisfy the customers. The reasons that affect the performance of the service are:

### 2.1 Scheduling

The task of creating a schedule, making the decision, arranging the task in First In First Out (FIFO) manner and for providing the resources based on the priority among the possibility of tasks is known as scheduling. The software that is used for handling scheduling process is known as scheduler. The scheduler is arranging the task in the sequence order. The scheduling mechanism is used for scheduling the jobs in the First In First Out (FIFO) manner for the full utilization of resources and for improving the performance of the system.

### 2.2 Oversubscription

In cloud resource allocation, oversubscription is meant for subscripting the cloud resources more that of the actual capacity of the environment. The subscription of customers for the cloud resources exceeds the available capacity of the cloud environment, it is called as oversubscription. Because of oversubscription the overload occurs. In some cases, the customers will not utilize the complete resources whatever they consumed. In this case, the resources are idle and waste. But for the functioning of the data centers, the high level of power and other resources are needed. The service providers spend a lot of amount for the maintenance of the data centers. When the resources are not utilized as reserved by the customers. Therefore, the service providers allow oversubscription for the complete utilization of resources but it leads the problem of overload.

### 2.3 Congestion

In cloud environment, congestion is the conditional level of excessive resource allocation in providing the resources to the customers. When oversubscribe the resources, the load is increased; and it is higher than the capacity of the cloud environment; so that, there is congestion occurs in providing the resources to the customers. Because of congestion, the resources cannot be provided as defined in the SLA. It affects the quality of service (QoS) metrics. The performance and efficiency in providing the resources will be poor; sometimes, it drops the provision of resources. Therefore, the resources are to be provided or transmitted again. The bandwidth will be wasted for retransmissions.

## 3.RELATED WORKS

Soumya Ranjan Jena et al.[1] discussed about three major load balancing algorithms to reduce the response time. They are Round robin, Active monitoring and Throttled. The Round robin algorithm is an efficient method when VMs are having one processor. But it is not possible in the system. The experiment shows that response time and data processing time are minimized by using active monitoring load balancing algorithm. Throttled load balancing algorithm gives the better result than the Round robin algorithm. Therefore the author concluded that Active monitoring load balancing algorithm is an efficient one than the other two algorithms. But Active monitoring is not the appropriate method for dynamic allocation of resources.

Mohit Kumar et al. [3] discussed the concept of load balancing in cloud computing to minimize the response time, the author followed two methods that are SLA aware load balancing and (Join Idle Queue) JIQ. This leads to balance the load of VMs by calculating the processing time and minimum queue length. The aim of this method is to reduce the average response time, execution time and the waiting time of tasks. But, the applied methods are suitable for static allocation of resources not for dynamic allocation of resources because it starts with the allocation of virtual machines to the tasks before scheduling the tasks in the workflow.

Jitendra Singh et al. [6] discussed about the response time dependency on broker service policy and the number of data centers. There are three broker service policies : Closest Data Center(CDC), Optimum Response Time(ORT) and Reconfigure Dynamically with Load(RDL) . The three broker service policies are evaluated; the Closest Data Center(CDC) provides the best performance with a minimum response time. The author concluded that the Closest Data Center(CDC) is the best service policy for better performance and the number of data centers to be increased to reduce the response time. This paper did not discuss about the technic for minimizing the workload of system.

G.Rajiv Ratnakar et al. [7] discussed about the load balancing algorithms to reduce the response time. They are Round robin and Throttled load balancing algorithms. The Round robin algorithm is an efficient method for static allocation of resources and Throttled load balancing algorithm gives the better result than the Round robin algorithm for dynamic allocation of resources. Cloud analyst is used as tool to evaluate the performance of the service in two ways: user base and data center. Equally spread current execution is a method to reduce the workload of the system. The experiment shows that response time is minimized by using Throttled load balancing algorithm in the dynamic scheme. Therefore the author concluded that Throttled load balancing algorithm is an efficient one for dynamic allocation of resources with minimum response time. In this paper, the partitioning method for dynamic allocation is to be implemented.

Maryam Houtinezhad et al. [9] discussed about particle swarm optimization algorithm for the minimization of cost and a method to calculate the penalty, the particle swarm optimization algorithm is combined with Meta heuristic algorithm and genetic algorithm for best performance and for the minimization of cost. A scheduling policy is provided to allocate the resources and to improve the performance of the system. This method will not be profitable for the users.

Vidhi Tailong el al. [12] discussed about modified optimize response time service broker policy (optimize response time) of cloud analyst; the author integrated the sorting and mapping technique to reduce the response time. The policy registers all the data centers in ascending order based on the response time; therefore it can map the

user bases based on the registered list of data centers. The round robin algorithm is used for the distribution of load among data centers; the modified optimize response time service broker policy shows better results with round robin algorithm.

Subhash. B. Malewar et al. [14] discussed the Effective Load Balance (ELB) algorithm to minimize the response time. It is implemented using the Cloud-Analyst simulator. The response time is evaluated from different data center at different region with user bases. The user is requesting VM from a same region or from different regions. The proposed ELB algorithm minimizes the response time considerably and complete job faster than other algorithms. This method directly following the allocation of virtual machines but it should follow the scheduling process first for the tasks which are in the workflow.

Ziqian Dong et al. [17] discussed about the formulation of task assignment to a data center as an integer programming optimization problem and demonstrated that the average response time of task is bounded with number of active servers. The author proposed a greedy task-scheduling algorithm, to reduce the consumption of power of a data center by using the Most Efficient Server First scheme. The MESF scheduling scheme schedules tasks, which are in the efficient servers of the required data center. This scheme minimizes the average task response time and minimizes the server-related energy expenses.

## 4.AN INVESTIGATION OF THE EXISTING SCHEDULING APPROACHES

An investigation of the existing scheduling approaches and the various scheduling parameters that are used by them, benefits and drawbacks are summarized in TABLE 1.

| | | | |
|---|---|---|---|
| Round Robin | The time of arrival and quantum of time | There is no deadline based task. A quantum of time is given for all tasks. | Before the process starts, the quantum time is to be calculated based on the number of tasks in the workflow |

TABLE 1
AN INVESTIGATION OF EXISTING SCHEDULING APPROACHES
VARIOUS SCHEDULING PARAMETERS BENEFITS AND DRAWBACKS

| Scheduling Approaches | Scheduling Parameters Focused | Benefits | Drawbacks |
|---|---|---|---|
| First Come First Serve | The time of arrival | Simple Approach. The process is carried out easily. | It considers arrival time only; The other parameters are not focused. |
| Max-Min | Processing time & expected completion time for big tasks | Better processing time for the allotted big tasks | The time is not balanced so that small tasks are in the workflow as deadline based tasks |
| Min-Min | Processing time & expected completion time for small tasks | Better processing time for the allotted small tasks | The time is not balanced so that big tasks are in the workflow as deadline based tasks |
| Priority based scheduling | Priority of the task or Priority of the user request or Expected execution time | The scheduling process is followed based on the priority. According to the priority, the decision making process for the application is developed. | The tasks are waiting for a long time to the assigned CPU because of infinite blocking. The complexity to be minimized. |
| Switching Algorithm | Processing time, high performance and balancing workload | Better processing time because the scheduling process is followed based on the workload of the system. | The cost is increased based on time. It is benefitable one to the provider not for the customers. |
| Improved cost based algorithm | Resource cost and processing time | Before starting the scheduling process, the parameters resource cost and the processing time are focused. | The other parameters are not considered for the quality of service. |

## 5.ISSUES AND CHALLENGES

1. When the proper scheduling process is not followed, the systems that cannot decide which process run at certain point in time.
2. When the subscription is increased, the workload is increased, so the problem of overload occurs and there will

be congestion in providing the service, therefore the response time will be maximized. Balancing the load is an issue in cloud computing environment.

3. When the congestion occurs, there will be bottle neck problem so the jobs will be in the workflow as deadline based jobs.
4. The SLA oriented resource allocation is a challenge for differentiating and distributing the resources and for satisfying the customers by providing the desired utility to the required customers.
5. The service quality parameters should be mentioned when define a SLA, the customers can evaluate the service by the defined parameters. The feedback mechanism is used for providing the quality of service; this is a way for encouraging the service request.
6. It is critical for designing a system which is supporting a fully Service-Oriented resource management in cloud computing environments.
7. To satisfy the customers, the providers proposed three users –centric objectives in the context of cloud computing service.

## 6.CONCLUSION

This paper discussed the issues improper scheduling, oversubscription, overload and congestion for the maximum level of response time in cloud environment. To provide the best quality of service (QOS), the scheduling and workload should be analyzed from a capacity planning and resource provisioning perspective, for minimizing the response time. Because of oversubscription the workload is increased and there will be congestion in providing the service, therefore the response time will be maximized. To minimize the response time the proper scheduling mechanism, congestion control mechanism and the proper tools to be applied to avoid the workload in cloud environment. The scheduling mechanism is the solution to avoid deadline based jobs. Moreover, the proper load balancing method is used for minimizing the response time, high performance and to provide the proper resource utilization and to avoid the overload. There are many approaches for doing the research in different levels. Deeper studies on minimizing the response time in cloud environment to deal with the issues improper scheduling, oversubscription, overload and congestion control related to the cloud environment can be focused for future work.

## REFERENCES

[1] Soumya Ranjan Jena and Zulfikhar Ahmad, "Response Time Minimization of Different Load Balancing Algorithms in Cloud Computing Environment", International Journal of Computer Applications (0975 – 8887), Volume 69– No.17, May 2013, http://research.ijcaonline.org/volume69/number17/pxc3888144.pdf

[2] Funmilade Faniyia, Rami Bahsoona, Georgios Theodoropoulosb, "A Dynamic Data-Driven Simulation Approach for Preventing Service Level Agreement Violations in Cloud Federation", International Conference on Computational Science, ICCS 2012, 1877-0509 © 2012 Published by Elsevier Ltd., doi: 10.1016/j.procs.2012.04.126

[3] Mohit Kumar, "Load Balancing Algorithm Using JIQ in Cloud Computing", International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN: 2349-2763, Issue 01, Volume 4 (January 2017), http://ijirae.com/volumes/Vol4/iss1/01.DCAE10087.pdf

[4] Ahmad Mosallanejad, Rodziah Atan, Masrah Azmi Murad, Rusli Abdullah, "A Hierarchical Self-Healing SLA for Cloud Computing", International Journal of Digital Information and Wireless Communications (IJDIWC) 4(1): 43-52, The Society of Digital Information and Wireless Communications, 2014 (ISSN:2225-658X)

[5] Lionel Eyraud-Dubois, Hubert Larcheveque, "Optimizing Resource allocation while handling SLA violations in Cloud Computing platforms", IPDPS - 27th IEEE International Parallel & Distributed Processing Symposium, May 2013, Boston, United States. 2013, <10.1109/IPDPS.2013.67>. <hal-00772846>

[6] Jitendra Singh, "Study of Response Time in Cloud Computing", I.J. Information Engineering and Electronic Business, 2014, 5, 36-43 Published Online October 2014 in MECS DOI: 10.5815/ijieeb.2014.05.06, http://www.mecs-press.org/ijieeb/ijieeb-v6-n5/IJIEEB-V6-N5-6.pdf

[7] G.Rajiv Ratnakar and CH.Madhu Babu, "Improved Load Balancing Model Based On Partitioning In Cloud Computing", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, IJCSMC, Vol. 3, Issue. 7, July 2014, pg.955 – 959, http://ijcsmc.com/docs/papers/July2014/V3I7201499a67.pdf

[8] S.Sujan, R.Kanniga Devi, "A Dynamic Scheduling Scheme for Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), ISSN: 2278 – 1323 Volume 4 Issue 3, March 2015

[9] Maryam Houtinezhad, Kerman Branch, Amid Khatibi Bardsiri And Bardsir Branch, "Minimizing Response Time for Scheduled Tasks Using the Improved Particle Swarm Optimization Algorithm in a Cloud Computing Environment", Service Technology Magazine, Issue XCI • July/August 2015, www.servicetechmag.com, http://servicetechmag.com/I91/0715-3

[10] Sahar Mohamed Musa, Adil Yousif, Mohammed Bakri Bashi, "SLA Violation Detection Mechanism Cloud Computing", http://www.ijcaonline.org/archives/volume133/number6/23788-2015907483 , "International Journal of Computer Applications (0975 – 8887) Volume 133 – No.6, January 2016

[11] Vincent C. Emeakarohaa, Marco A.S. Netto, Rodrigo N. Calheiros, Ivona Brandic, Rajkumar Buyyac, César A.F. De Rose, "Towards autonomic detection of SLA violations in Cloud infrastructures" , Future Generation Computer Systems, doi:10.1016/j.future.2011.08.018, www.cloudbus.org/papers/AutonomicSLA-Cloud-FGCS2012.pdf

[12] Vidhi Tailong and Vivek Dimri, "Load Balancing in Cloud Computing Using Modified Optimize Response Time", International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSE _Volume 6, Issue 5, May 2016, pp. 552-557 ISSN: 2277 128X, www.ijarcsse.com, http://ijarcsse.com/docs/papers/Volume_6/5_May2016/V6I5-0273.pdf

[13] Marc Eduard Frîncu, "Scheduling highly available applications on cloud environments", Future Generation Computer System(2012), doi:10.1016/j.future.2012.05.017

[14] Subhash. B. Malewar and Prof-Deepak Kapgate, "Effective Virtual Machine Scheduling in Cloud Computing", International Journal For Research In Emerging Science And Technology, Volume-2, Issue-5, May-2015, E-ISSN: 2349-7610, http://ijrest.net/downloads/volume-2/issue-5/pid-ijrest-25201579.pdf

[15] Hossein Morshedlou, Mohammad Reza Meybodi, "Decreasing Impact of SLA Violations: A Proactive Resource Allocation Approach for Cloud Computing Environments" IEEE Transactions On Cloud Computing, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.699.9120, VOL. 2, NO. 2, APRIL- JUNE 2014, 2168-7161 _ 2014 IEEE.

[16] P.Premkumar, Dr.D.Shanthi, "An Efficient Dynamic Data Violation Checking Technique For Data Integrity Assurance In Cloud Computing", ISSN (Online) : 2319 – 8753, ISSN (Print) : 2347 – 6710, International Conference on Innovations in Engineering and Technology (ICIET'14) , Volume 3, Special Issue 3, On 21st & 22nd March 2014

[17] Ziqian Dong, Ning Liu and Roberto Rojas-Cessa, "Greedy Scheduling Of Tasks With Time Constraints For Energy-Efficient Cloud-Computing Data Centers", Journal of Cloud Computing: Advances, Systems and Applications (2015) 4:5, © 2015; licensee Springer, DOI 10.1186/s13677-015-0031-y, https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-015-0031-y

[18] Manpreet Kaur, Jaspreet Kaur & Sahil Vashist, "Autonomic Brokering For Minimization Of Service Level Agreement Violations In Cloud Computing", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), ISSN(P): 2249-6831; ISSN(E): 2249-7943, Vol. 4, Issue 5, Oct 2014, 23-30, © TJPRC Pvt. Ltd.

[19] K. Vaitheki & S. Urmela, "A SLA violation reduction technique in Cloud by Resource Rescheduling Algorithm (RRA)", International Journal of Computer Application and Engineering Technology, Volume 3-Issue 3, ISSN : 2277 7962, July 2014. Pp. 217-224, www.ijcaet.net

[20] S Anithakumari, K Chandrasekaran, "Autonomic Cloud Computing: Autonomic Properties Embedded in Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN : 2277 128X , Volume 5, Issue 4, 2015. April- 2015, pp. 979-991

[21] Michael Maurer, Ivona Brandic and Rizos Sakellariou," Enacting SLAs in Clouds Using Rules", http://www.cs.man.ac.uk/~rizos/papers/europar11b.pdf, Euro-Par 2011, LNCS 6852, Part I, pp. 455–466, 2011. c_Springer-Verlag Berlin Heidelberg 2011

[22] A. V. Dastjerdi and R. Buyya, (2012). An Autonomous Reliability-Aware Negotiation Strategy for Cloud Computing Environments. In Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid), Ottawa, Canada.

[23] S. Anithakumari and K. Chandra Sekaran, "Autonomic SLA Management in Cloud Computing Services", http://link.springer.com/chapter/10.1007%2F978-3-642-54525-2_13, Springer Berlin Heidelberg, Series ISSN : 1865-092, Online ISBN : 978-3-642-54525-2, Print ISBN : 978-3-642-54524-5, DOI : 10.1007/978-3- 642-54525-2_13, pp 151-159

[24] Rajkumar Kettimuthu, "Type- and Workload-Aware Scheduling of Large-Scale Wide-Area Data Transfers Dissertation", Ohio State University, 2015. OhioLINK Electronic Theses and Dissertations Center. https://etd.ohiolink.edu/!etd.send_file?accession=osu1437747493&disposition=inline, 29 Oct 2016.

[25] Rajkumar Rajavel, Mala, "Achieving Service Level Agreement in Cloud Environment using Job Prioritization in Hierarchical Scheduling", Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, Pages : 547-554, Publisher : Springer Berlin Heidelberg, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=t_bhV60AAAAJ&citation_for_view=t_bhV60AAAAJ:d1gkVwhDpl0C

[26] Liang Zhao, Sherif Sakr, Anna Liu, "A Framework for Consumer-Centric SLA Management of Cloud-Hosted Databases", https://www.computer.org/csdl/trans/sc/preprint/06461875.pdf, 1939-1374/13/$31.00@ 2013 IEEE

[27] Hadi Goudarzi, Mohammad Ghasemazar, and Massoud Pedram , "SLA-based Optimization of Power and Migration Cost in Cloud Computing", http://ieeexplore.ieee.org/document/6217419/, Print ISBN : 978-1-4673-1395-7, CD-ROM ISBN: 978-0-7695-4691-9, INSPEC Accession Number : 12803853, Publisher : IEEE, DOI : 10.1109/CCGrid.2012.112, Date of Conference: 13-16 May 2012

[1]