

A Two Stage Task Scheduler for Effective Load Optimization in Cloud – FoG Architectures

J. Samuel Manoharan

Professor, Department of Electronics & Communication Engineering, Sir Isaac Newton College of Engineering and Technology, Nagapattinam, India

E-mail: drjasm1530@ieee.org

Abstract

In recent times, computing technologies have moved over to a new dimension with the advent of cloud platforms which provide seamless rendering of required services to consumers either in static or dynamic state. In addition, the nature of data being handled in today's scenario has also become sophisticated as mostly real time data acquisition systems equipped with High-Definition capture (HD) have become common. Lately, cloud systems have also become prone to computing overheads owing to huge volume of data being imparted on them especially in real time applications. To assist and simplify the computational complexity of cloud systems, FoG platforms are being integrated into cloud interfaces to streamline and provide computing at the edge nodes rather at the cloud core processors, thus accounting for reduction of load overhead on cloud core processors. This research paper proposes a Two Stage Load Optimizer (TSLO) implemented as a double stage optimizer with one being deployed at FoG level and the other at the Cloud level. The computational complexity analysis is extensively done and compared with existing benchmark methods and superior performance of the suggested method is observed and reported.

Keywords: Cloud Computing, Task scheduler, FoG Computing, K-Means Clustering, Ada Boost Classifier



1. Introduction

Data technologies have experienced untold revolutions due to advent of state-of-the-art data acquisition technologies, cutting edge processing methodologies and storage mechanisms. In reality, amongst fast emerging computing technologies, there is a high degree of competition amongst manufacturers and service providers to provide high quality service to consumers based on their needs and requirements. The mentality of consumers in recent times is to have seamless access of data on their handheld and portable gadgets, irrespective of time and geographic locations. Evolution of communication technologies in the form of fourth generation, fifth generation and the sixth-generation systems which are at concluding stages of research, put a great deal of attractiveness for access of data by consumers. However, provision of services at high data rates and quality, as simple as it may sound, is quite sophisticated and requires cutting edge technologies. As a result of rigorous research in the past two decades, the conventional one on one processing and service provision technologies involving manual labor, intensive resource allocation mechanisms and exorbitant time-delays, cloud computing platforms have emerged to be the key players in the common ground of data acquisition, processing and service provision to consumers. The service provision is normally graded in terms of Quality of Service (QoS) which is a collection of several attributes like throughput, latency, transmission rate, customer satisfaction, optimal utilization of resources etc.

A simple cloud computing platform depicted in figure 1 is where the client infrastructure may be fixed or mobile as in the case of mobiles, laptops, iPad, virtual assistants etc. The requirement of data or certain service by the consumer is passed over into the cloud interface, which is analyzed in-depth for its requirement, the type of resource requested etc. and appropriately a job scheduler is assigned to attend to the client request. This is quite a straightforward approach in case of a single consumer-cloud platform or a few users. However, in case of large number of

users with different types of requests with varying levels of sophistication in their demands, conventional cloud systems must invoke concepts of optimality, streamlining, data grouping etc. to handle the bulk requests. This becomes an added overhead in case of huge organizations and in case of multi-organization deployment scenario where multiple organizations have their data processing being vested with cloud providers. Advancements in data acquisition, advent of Internet of Things (IoTs) which capture a huge volume of bulk data or big data from various sources (heterogeneous) pushes the cloud cores to their maximum limits of operation. In order to address this computational complexity issue of handling large streams of heterogeneous data, FoG or Edge Computing [1] has been on the rise in the past decade. A typical FoG based Cloud platform is depicted in figure 2.

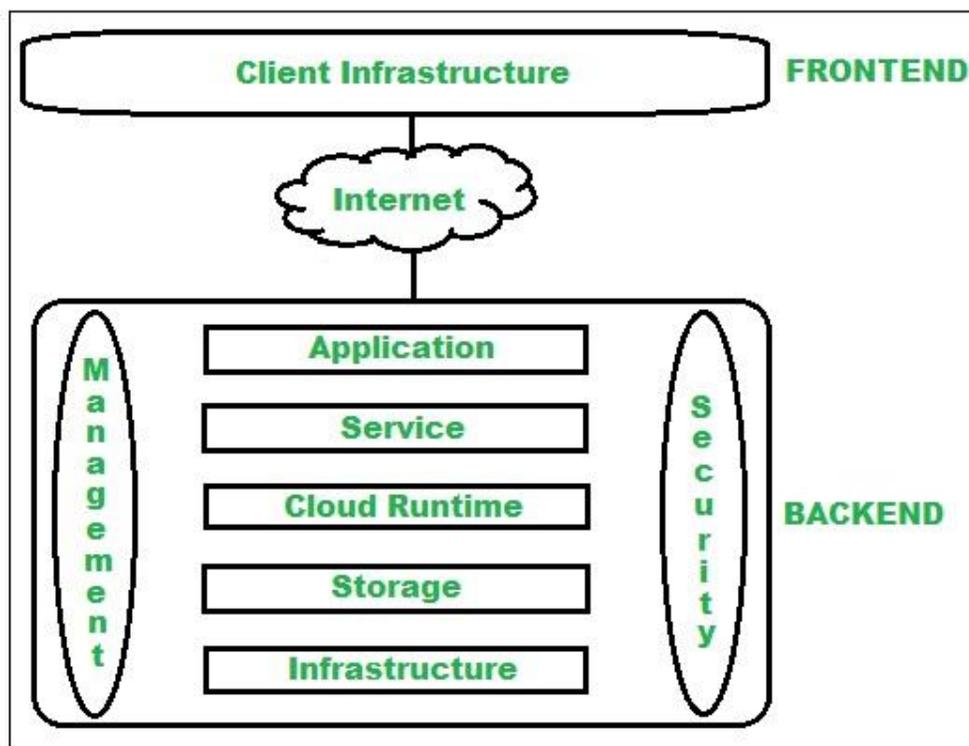


Figure 1. A typical cloud computing architecture [1]

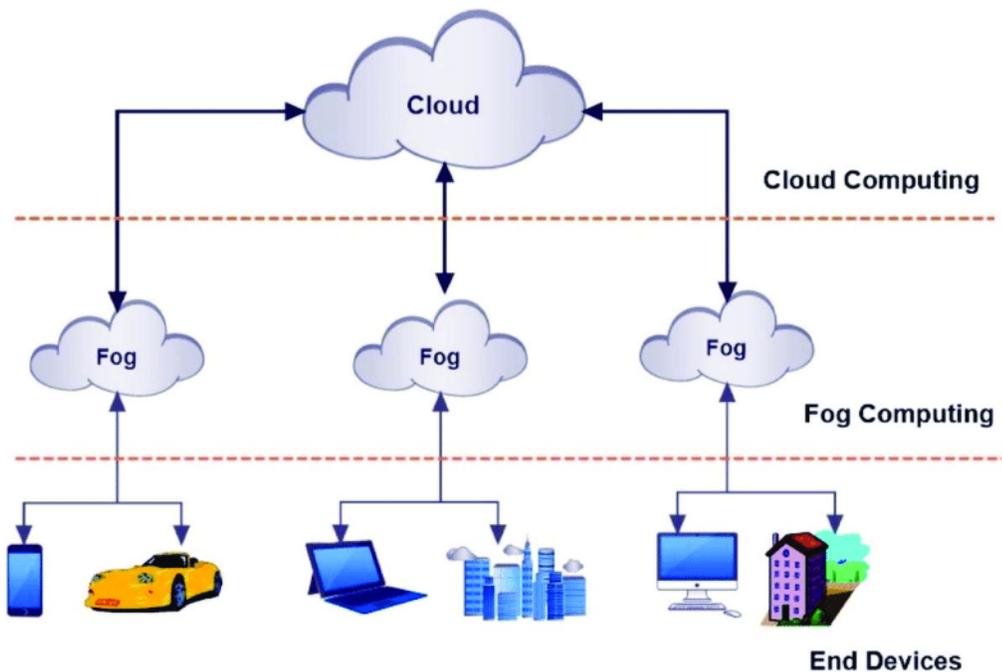


Figure 2. Illustration of FoG based Cloud Architecture [1]

Figure 2 provides a simplified picture of FoG assisted cloud computing wherein the requirements of the consumers or users are handled by the FoG nodes or the Edge nodes rather than going in for the Cloud Core Level (CCL). However, in places where processing involving the resources stored in cloud storage is required, the FoG nodes need to communicate to the CCL to provide the request/demand placed by the consumers. A FoG assisted Cloud platform ensures a decentralization of processing to a certain level thus reducing the burden on the CCL to a great extent.

However, in case of large data streams from multiple sources arriving at the FoG nodes, a certain level of regulation of data [2] should be ensured by FoG nodes failing which, the complexity of the FoG nodes would increase exponentially thus causing the failure of the node or the objective for which it was intended to be used for. After a certain regulation of data has been

done at the FoG layer, further processing at the cloud level should also be regulated for optimal assignment of resources in order to enhance the QoS of the overall network [3]. This is alternately termed as the task scheduling approach. A typical task scheduling model is shown in figure 3.

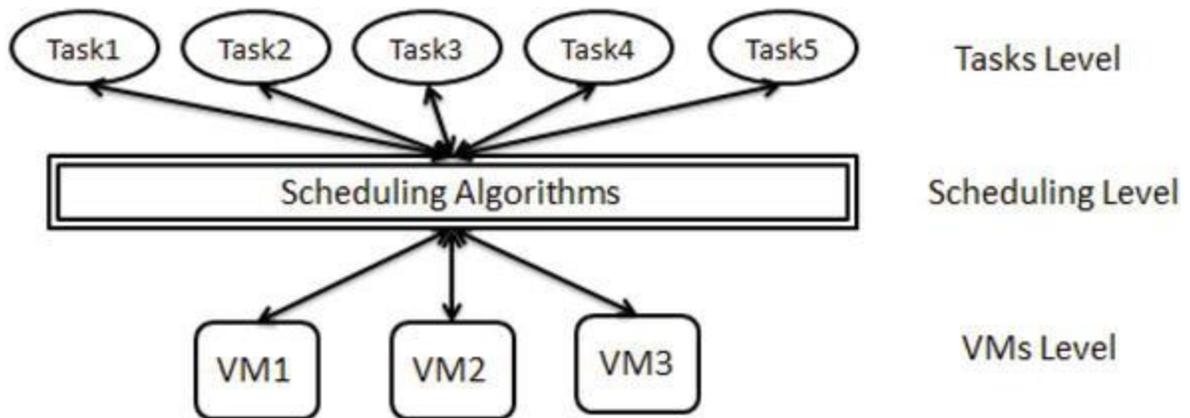


Figure 3. Illustration of a simple task scheduler scheme [3]

This has been taken as the problem definition in this research work and a novel two stage load balancing methodology has been proposed with one being implemented at the FoG Layer of Computing (FLC) and other at the Cloud Layer of Computing (CLC).

The rest of the paper is organized into a brief survey of literature in section 2 followed by the proposed two-stage model for load balancing in section 3. Experimentations and observations have been discussed in section 4 followed by the concluding remarks of the paper in section 5.

2. Related Work

Several works are found in the literature related to cloud implementations. Recent and most relevant works to the proposed objective have been explored and summarized in this section. One

of the most challenging and researched topics in cloud computing pertains to the problem of load balancing [4-10] of tasks or jobs to minimize the overhead on the cloud framework.

Prediction based models [11, 12] have been implemented for estimating the load requirement or demand based on previous and current data traffic pattern analysis methods. Concepts of Machine learning have been effectively utilized. However, prediction-based methods are unstable for such huge data volume based IoT and FoG based cloud architectures thus making them unsuitable candidates. Round robin scheduling [13, 14] is the simplest yet effective method of scheduling which works on the basis of time slot formulation. A circular linked list is formulated, and tasks are assigned on the basis of time slots to the processors or the VMs. This method is effectual for a low-level intensity job request flow-based applications and not suitable for large voluminous data. A lot of opportunistic algorithms are available in the literature [15-18] wherein the tasks are just assigned to idle node points or VMs irrespective of analyzing their computation capability, available resources or latency. Hence, it is more of a random kind of assignment and generally results in poor QoS in terms of load balancing especially in cases of large heterogeneous data. In a heterogeneous situation, different requests from different users may require different computational requirements and overheads thereby alleviating the research challenge associated with it.

Utilization of Map reduce Hadoop [19, 20] is yet another effective formulation for load balancing in cloud platforms. A simple Hadoop formulation is depicted in figure 3. A cluster-based approach is used to group and separate the incoming data into partitions or clusters based on similarity measures and map them onto the available set of VMs. However, the computational complexity increases when applied for FoG based cloud platforms. The concept of dynamicity has been accounted for in the form of a hybrid algorithm [21-23] where a dynamic balancing parameter has been introduced to minimize the access time. Reduced waiting time is the essence of the

experimental result. The processor capability and availability of resources have been taken into account. However, an integrated FoG based cloud platform has not been tested in this work.

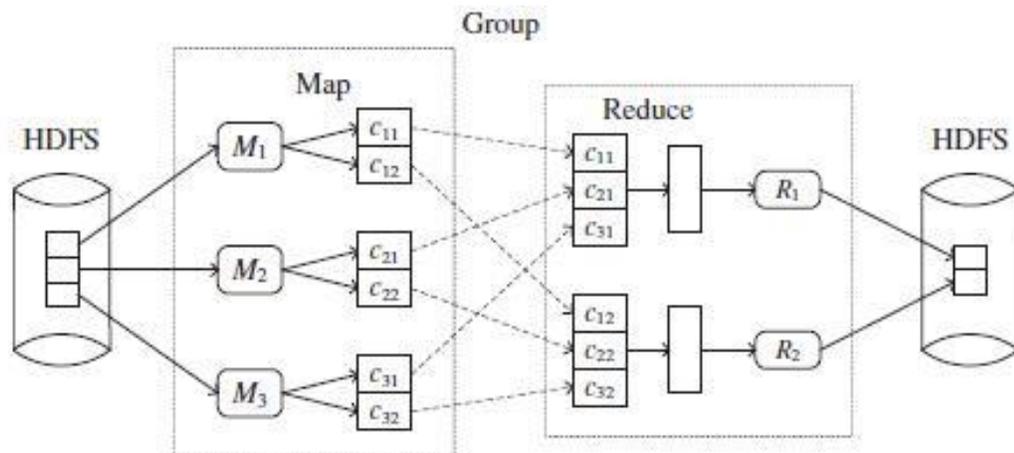


Figure 3. Map Reduce Model for Cloud computing [21]

In recent times, Meta-heuristics [24] have been playing a key role in solving the load optimization problems. These are a class of algorithms that work on optimization, towards a target objective based on a set of constraints or bounds. A class of meta-heuristics known as nature inspired algorithms, are based on naturally occurring phenomena and mimic real time applications due to which their application normally yields accurate and optimal results. However, the computational complexity is to be handled in a systematic manner. Many such implementations like the genetic algorithm-based optimizer [25], the ant colony-based optimization algorithm [26], particle swarm optimization-based resource allocation models [27] etc. have been observed.

Concepts of learning and deep learning have also been invoked in the literature starting from a simple, yet effective ANN based optimizer [28] for learning the pattern of traffic and implementing scalability in the allocation process. Genetic concepts have been integrated for obtaining optimization in the resource allocation process [29, 30]. As a general conclusion, it could

be stated that, critical research works have been contributed towards energy aware load balancing in cloud platforms. Most of them have reported optimal performances but however not tested in an integrated FoG assisted cloud network for IoT applications which has aided to be the motivation behind the work proposed in this research paper.

3. Proposed Work

A two-stage load optimizer/balancer forms the essence of the proposed work. The first level of load optimization is done by adopting the methodology of data grouping where data from several sources of the IoT are grouped based on some sort of similarity. Clustering is a well-known grouping strategy and a K-Means clustering methodology has been adopted in the proposed work at the FLC layer.

Given a stream of heterogenous data from various sources $S = \{S_1, S_2, S_3, \dots, S_k\}$ where $k \in \mathbb{R}$ reflects the maximum number of routes or IoT data input sources. Assuming that there are $F = \{F_1, F_2, S_3, \dots, F_l\}$, where $F_l \leq S_k$, the objective of the proposed K-Means would be to map the set of all $S \rightarrow F$.

It is to be noted in the above formulation that, the value of l in F_l is very negligible as compared to k in S_k . Hence, the effectiveness of the load balancer lies in mapping the data from a vast set of IoT sources onto a limited number of FoG nodes so that all data are accounted for and serviced in the least possible time. Accordingly, the K-Means algorithm for the proposed work is formulated as

Input: Data Stream $S\{ \}$, Set of FoG nodes $F\{ \}$

Output: Cluster Group $C_G, S \rightarrow F \in \mathbb{R}$ with $S[S_k == 0]$

```
begin
{
    initialize F, S ∈ R, number of clusters M

    while G ≠ 0
    {
        for i = 1: Pj

        compute CG ∈ C → P * computing closest centroid

        CG ← Pj

        update and set CG as new center
    }
}

If CG does not change, then terminate

end
```

The algorithm proceeds by analyzing the set of input sources from several sources which may involve an IoT application. The incoming rate of information is analyzed to assign the FoG nodes accordingly. Euclidean distance is taken as the similarity measure to group the data stream into clusters and unassigned data points are left as such. The clustered groups are segregated into two instance lists I_A and I_B where the former denotes the requests requiring resource allocation from Cloud Core (CC) while the latter reflects the data points which do not require the utilization

of CC. The former set of instances are forwarded to CLC while the latter set of instances are processed and disposed-off by the edge nodes. The process is repeated until all the data points are accounted for. The instance segregation algorithm is reflected as below

Input: Cluster Group C_G ,

Output: Clustered Instances I_A and I_B

```
begin
{
    initialize  $I_A$  and  $I_B \in R$ 
    while all data points  $d_i \in C_G == 0$ 
    {
        for  $i = 1:k$ 
        compute  $res_{call}(d(k))$ 
        if  $res_{call}(d(k)) == 1$ 
        then  $I_A \leftarrow d(k)$ 
        else  $I_B \leftarrow d(k)$ 
    }
} end
```

An Adaboost classifier is utilized in the CLC level to formulate and implement an optimal resource allocation protocol. The clustered group I_A which has been forwarded by the FLC layer is received as the input at CC. The attributes of the data points are studied by the classifier model and a pattern map is generated. This pattern map corresponds to the resource utilization required for each clustered group in the instance I_A . Based on the pattern map, the Adaboost classifier assigns the priority to the data points based on which the Virtual Machines (VMs) are allocated to accomplish the required objective. A simple priority-based job scheduler is implemented to act on the output of the Adaboost classifier to resource allocation.

4. Experimental Work

A vast set of experimentations have been conducted to test the efficiency of the proposed two stage load balancing algorithm for the proposed FoG based Cloud computing architecture. Cloud Sim Toolkit has been effectively utilized to formulate the Cloud and Fog architecture model. The Cloud Sim Toolkit works on a Java back-end platform in which the cloud and FoG nodes are initialized. A snapshot of the initialization carried out on the Cloud Sim Toolkit is depicted in figure 4.

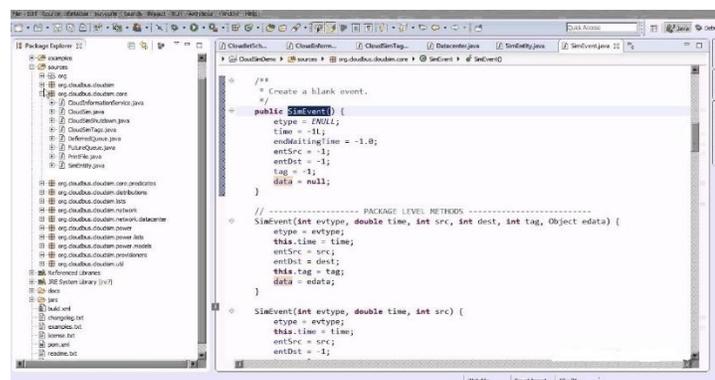


Figure 4. Initialization of Cloud and FoG node in Cloud Sim Toolkit

The initialization of various attributes related to the architecture are listed in the table shown below.

Table 1. Simulation Settings – Proposed Two Stage Load Optimizer (TSLO)

Attribute	Value
No. of Tasks	300
Length/Task	7000
Data set size (Big Data)	4000
No. of VMs	15
Bandwidth	1000
Memory	512KB
Storage	1TB
No. of Fog Nodes	10
Bandwidth	750
Storage	1GB
Memory	256KB
Existing load percentage	32%
Number of requests	48
Virtual machine count	12
Experiment duration	18400s

The analysis has been done primarily based on metrics like waiting time, data center utilization analysis and response time. The above listed metrics effectively visualize the efficiency of the proposed two stage load balancing scheduler. The performances have been compared with recent benchmark methods which work based on opportunistic and dynamic parameter-based load balancing methods.

Table 2 depicts the execution time analysis of the virtual machines in the CLC based on varying utilization of VMs depending on the number of requests placed from the FLC.

Table 2. Comparative performance of Execution time for varying VM

No. of VMs	Execution time (ms)		
	Opportunistic based	Dynamic based	Proposed TSLO
2	110	96	79
4	120	166	94
6	210	205	140
8	400	388	202
10	790	670	410
12	1024	1004	640

Waiting time analysis is also an essential metric which helps visualize the availability of VM or resource which is demanded by the consumer through the FLC to the CLC. Figure 5 reflects the waiting time analysis which is observed to be minimal in case of the proposed TSLO algorithm.

This is primarily due to the quick allocation of resources of VMs from the CLC. Also, the filtering out of instances at the FLC which do not require the need of VMs or the resources helps to drastically reduce the waiting time analysis. In this way, more critical applications or data points are assigned to VMs than the non-critical ones.

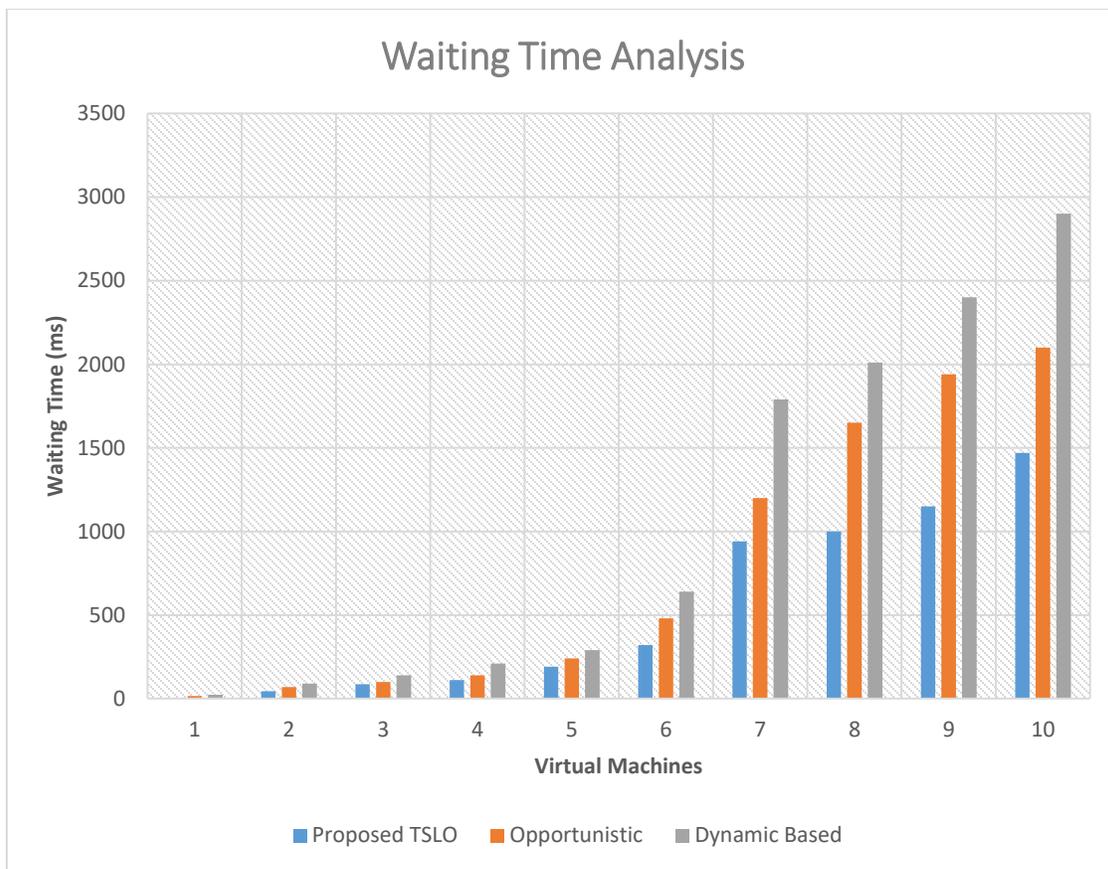


Figure 5. Waiting Time analysis – Comparative analysis

Another important analysis done is the data center utilization which constitutes to an energy aware and optimal scheduling approach. A comparative analysis is projected in the figure shown below.

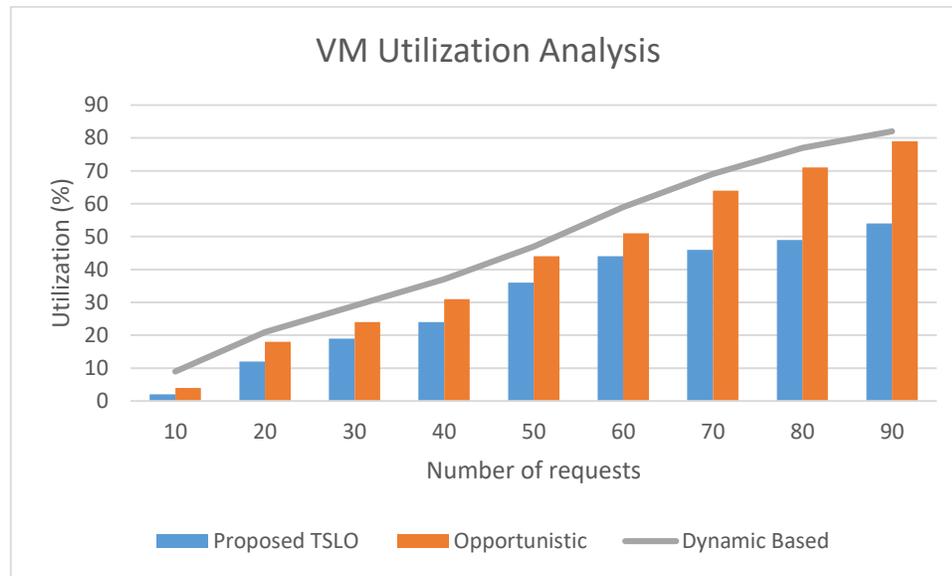


Figure 6. Comparative analysis of data center utilization

As observed in figure 6, a minimal utilization of data center or the resource has been observed in the proposed TLSO algorithm as compared to the opportunistic and dynamic allocation-based models. This is once again attributed towards filtering out unwanted resource requirements present in the instances which do not require resources from the CC.

5. Conclusion

Most of the computing technologies and organizations have migrated towards cloud computing platforms which are a way of providing effective services to the consumers on demand. This is achieved through a set of strong resource allocation and job schedulers in the core computing level of the cloud. However, with ever increasing data streams, need for reducing the computational complexity of the standalone cloud is strongly felt thus motivating the utilization of a FoG assisted cloud platform. A two stage load balancing approach has been proposed and

implemented in this research work to effectively handle the problem of task scheduling. The proposed two stage load optimization model has been compared with benchmark methods in the form of opportunistic scheduler and dynamic allocation-based scheduler and optimal performance has been observed and recorded over critical metrics like computational wait time, resource utilization analysis etc. With increasing data streams, more convergence is required in a short time possible which makes evolutionary algorithms to be probable candidates for integration with existing systems which can be thought as a future scope of the research work.

References

- [1] Atlam, Hany & Walters, Robert & Wills, Gary. (2018). Fog Computing and the Internet of Things: A Review. *Big Data and Cognitive Computing*. 2. 10.3390/bdcc2020010.
- [2] Shakya, Subarana. 2019. An efficient security framework for data migration in a cloud computing environment. *Journal of Artificial Intelligence*. 1(1): 45-53.
- [3] Pandian, A. Pasumpon, and S. Smys. 2020. Effective Fragmentation Minimization by Cloud Enabled Back Up Storage. *Journal of Ubiquitous Computing and Communication Technologies (UCCT)* 2(1): 1-9
- [4] Pandian, M. Durai. 2019. Survey on Virtual Load Balancing Architectures in Mobile Cloud." *IRO Journal on Sustainable Wireless Systems*. 1(3): 161-175.
- [5] Pradhan A., Bisoy S. K., Mallick P. K. (2020). Load Balancing in Cloud Computing: Survey. In: Sharma R., Mishra M., Nayak J., Naik B., Pelusi D. (eds) *Innovation in Electrical Power Engineering, Communication, and Computing Technology. Lecture Notes in Electrical Engineering*, vol 630. Springer, Singapore. https://doi.org/10.1007/978-981-15-2305-2_8.
- [6] Sungheetha, Akey, and Rajesh Sharma. "Real Time Monitoring and Fire Detection using Internet of Things and Cloud based Drones." *Journal of Soft Computing Paradigm (JSCP)* 2, no. 03 (2020): 168-174.

- [7] Einollah Jafarnejad Ghomi, Amir Masoud Rahmani, Nooruldeen Nasih Qader. (2017). Load-balancing algorithms in cloud computing: A survey, *Journal of Network and Computer Applications*, 88: 50 – 71.
- [8] Joe, C. Vijesh, and Jennifer S. Raj. "Deniable Authentication Encryption for Privacy Protection using Blockchain." *Journal of Artificial Intelligence and Capsule Networks* 3, no. 3 (2021): 259-271.
- [9] Afzal,Kavitha. (2019). Load balancing in cloud computing – A hierarchical taxonomical classification. *Journal of Cloud Computing*, 8(22): 1-8.
- [10] Bhalaji, N. (2021). Cloud Load Estimation with Deep Logarithmic Network for Workload and Time Series Optimization. *Journal of Soft Computing Paradigm*. 3(3): 234-248.
- [11] Vignesh Joshi. (2019). Load Balancing Algorithms in Cloud Computing. *International Journal of Research in Engineering and Innovation*, 3: 530 – 532.
- [12] Smys, S., and Haoxiang Wang. "Security Enhancement in Smart Vehicle Using Blockchain-based Architectural Framework." *Journal of Artificial Intelligence* 3, no. 02 (2021): 90-100.
- [13] Singh, P., Baaga, P., & Gupta, S. (2016). Assorted Load Balancing Algorithms in Cloud Computing: A Survey. *International Journal of Computer Applications*, 143(7): 34-40.
- [14] Mugunthan, S. R. "Soft computing based autonomous low rate DDOS attack detection and security for cloud computing." *J. Soft Comput. Paradig.(JSCP)* 1, no. 02 (2019): 80-90.
- [15] Madni, S. H. H., Latiff, M. S. A., & Coulibaly, Y. (2016). Resource scheduling for infrastructure as a service (IaaS) in cloud computing: Challenges and opportunities. *Journal of Network and Computer Applications*, 68:173-200.
- [16] Patil, Prachu J., Ritika V. Zalke, Kalyani R. Tumasare, Bhavana A. Shiwankar, Shivani R. Singh, and Shailesh Sakhare. "IoT Protocol for Accident Spotting with Medical Facility." *Journal of Artificial Intelligence* 3, no. 02 (2021): 140-150.

- [17] Almezeini, N. and Hafez, A. An Enhanced Workflow Scheduling Algorithm in Cloud Computing. In Proceedings of the 6th International Conference on Cloud Computing and Services Science (CLOSER 2016) - Volume 2, pages 67-73.
- [18] Shakya, Subarna. "A Self Monitoring and Analyzing System for Solar Power Station using IoT and Data Mining Algorithms." *Journal of Soft Computing Paradigm* 3, no. 2 (2021): 96-109.
- [19] Khan M, Jin Y, Li M, Xiang Y, Jiang C. (2016). Hadoop performance modeling for job estimation and resource provisioning. *IEEE Trans Parallel Distribution Systems*. 27(2):441–54.
- [20] Jian Yang, Zhihui Lu, Nini Wang, Jie Wu, Patrick C., Hung, K. (2017). Multi-policy-aware MapReduce resource allocation and scheduling for smart computing cluster” *Journal of Systems Architecture*, 80: 17-29.
- [21] Rai, S., Sagar, N., & Sahu, R. (2017). An Efficient Distributed Dynamic Load Balancing Method based on Hybrid Approach in Cloud Computing. *International Journal of Computer Applications*, 169(9): 16-21.
- [22] Yaser Jararweh, Manar Bani Issa, Mustafa Daraghme, Mahmoud Al-Ayyoub, Mohammad A. Alsmirat, (2018). Energy efficient dynamic resource management in cloud computing based on logistic regression model and median absolute deviation. *Sustainable Computing: Informatics and Systems*, 19: 262-274
- [23] Tan Xiaoying, Huang Dan, Guo Yuchun, Chen Changjia. (2017) Dynamic resource allocation in cloud download service” *The Journal of China Universities of Posts and Telecommunications*, 24(5): 53-59.
- [24] Kaur, S.; Sengupta, J. Load Balancing using Improved Genetic Algorithm (IGA) in Cloud computing. *International Journal of Advanced Research in Computer Engineering and Technology*. 6(8): 1229 – 1233.

- [25] Kalra, M. & Singh, S. 2015. A Review of Metaheuristic Scheduling Techniques in Cloud Computing. *Egyptian Informatics Journal*. 16(3): 275 – 295.
- [26] Kumar, A.S.; Venkatesan, M. (2019). Multi-Objective Task Scheduling Using Hybrid Genetic-ant Colony Optimization Algorithm in Cloud Environment. *Wireless Personal Communication*. 107: 1835–1848.
- [27] Najme Mansouri, Behnam Mohammad Hasani Zade, Mohammad Masoud Javidi. (2019) Hybrid task scheduling strategy for cloud computing by modified particle swarm optimization and fuzzy theory” *Computers & Industrial Engineering*, 130: 597-633.
- [28] Mohan Sharma, Ritu Garg. (2020). An artificial neural network-based approach for energy efficient task scheduling in cloud data centers. *Sustainable computing: Informatics and Systems*. 26. <https://doi.org/10.1016/j.suscom.2020.100373>.
- [29] Karl Mason, Martin Duggan, Enda Barrett, Jim Duggan, Enda Howley. (2018). Predicting host CPU utilization in the cloud using evolutionary neural networks. *Future Generation Computer Systems*, 86: 162-173.
- [30] Yongnan Zhang, Yonghua Zhou. (2018). Distributed coordination control of traffic network flow using adaptive genetic algorithm based on cloud computing. *Journal of Network and Computer Applications*. 119: 110-120.

Author’s biography

J. Samuel Manoharan is a professor in the Department of Electronics and Communication Engineering at Sir Isaac Newton College of Engineering and Technology, India. His area of research includes Digital Image and Signal Processing, Data Security and Cryptography, Embedded Systems, Biomedical Instrumentation, Artificial Intelligence, Robotics, Deep Learning, Cognitive Science, Ad-hoc Networks, Artificial Neural Network, Evolutionary Computing, Speech Recognition and Autonomous Systems.