A Novel cloud Resource Allocation and load balancing framework using non-linear optimization constraints

Siva Rama Krishna*, Dr. Mohammed Ali Hussain#
*Research Scholar, Dept. of CSE, Shri Venkateshwara University, Uttar Pradesh, India.
#Professor, Dept. of CSE, Shri Venkateshwara University, Uttar Pradesh, India.

Abstract— As the size of the cloud computing resources and services increases, it is difficult to handle load balancing due to computational cost and time. Since, most of the cloud service providers have their own type, type and price policies for computing resources, including other service features. Since, computational time and memory of the existing cloud scheduling models are not efficient in realtime cloud environment .The load balance between cloud resources ensures an efficient utilization of the physical infrastructure while minimizing runtime However, the main problem to the cloud service provider's is optimizing cloud service parameters such as reliability, flexibility, time limits and the task refusal rate. In order to overcome the cloud resource allocation and scheduling problems a novel cloud resource allocation and load balancer framework is used to improve the cloud allocation issues. Experimental results proved that the present load-balancing model has better performance than the traditional load balancing approaches on various cloud resources.

Keywords—Cloud resource allocation, load balancer, non-linear optimization.

I. INTRODUCTION

Now-a-days, scheduling of computing resources has become the major concern of different research scientists. It has the basic objective of reducing the task completion time significantly. In case of supercomputers, multi-processor scheduling involves different numbers of parallel processors having equivalent capacity. Apart from this, the data source is required to be centralized and interlinked with the help of a high speed channel among various processors. In the above scenario, the activities can transfer messages easily and more quickly. Computer networks involve clusters of homogeneous computers in order to behave just like a multiprocessor computer having distributed data sources. Along with the latest advancements of computer networks, the connecting links in between various computing entities have become faster. Here we can mention that, the latest applications require extended bandwidth, large storage and exchange of huge volumes of data. There are two important applications such as, multimedia and e-Science hose require huge volume of data. It is very much necessary to achieve better performance and quality of service. Now-a-days, the popularity of customized, high quality products along with quick delivery is forcing different organizations to upgrade their traditional production process. Hence, latest management and control systems can be implemented in order to achieve better efficiency, robustness, responsiveness, agility and re-configurability. The enhancement of sustainability and maintainability of manufacturing processes has become the prime concern of various researchers presently. Therefore, implementation of cloud services is very much beneficial. The process of digital transformation or digitization can be defined as a specific kind of process in which the interaction among physical and informal entities exists. The complete process of digital manufacturing can also be defined as a combination of supply, production and delivery within a networked organization. Presently, cloud platforms are considered as the most cost efficient platforms and these are suitable for numbers of different real world applications just like engineering and scientific applications. These above said applications mostly involve numbers of different processes or tasks in order to construct a workflow. These workflows are represented by Directed-Acyclic Graph. In a workflow diagram, different tasks are interlinked through directed edges in order to represent the data dependency among various tasks. These constraints are known as precedence constraints. Every individual task is executed by considering the starting data inserted through workflow or data inserted through the parent tasks. Basically, tasks of a particular workflow are mostly scheduled in nature. These tasks are usually executed in the distributed form across multiple processing elements without violating precedence constraints.

Since last decade, the media streaming services have become more popular on internet. There exist very high demands of dynamic videos from all over the world. As compared to the conventional large server clusters, geo-distributed clouds are more scalable and feasible in nature. Hence, geo-distributed clouds are considered as the perfect solution. The major objective of cloud is to dynamically composing and optimizing the necessary services at lower costs which is not at all possible in case of dedicated servers. Most of the media service providers implemented cloud computing as a large-scale content distribution infrastructure because of the elasticity in case of dynamic resource management process. The service providers usually set up various data centers at different places. The media streaming system has the responsibility to provide better services near to the actual customers. In case of static contents, content delivery networks are the better and feasible option as compared to cloud. Most of the latest CDN follow the pay-as-you-go pricing schemes. Hence, the service provider is only required to pay for whatever they have downloaded. But in case of dynamic contents, cloud based systems is the best option. If the contents are both static and dynamic, in that case, there is a necessity of hybrid model.

Infrastructure as a Service (IaaS) clouds have the responsibility to provide the computing ability as a service to the customers. It also follows the basic rule of pay-per-use. Hence we can mention here that, very complicated and large execution environments can be deployed on cloud without establishing their own infrastructure. Presently, clouds are present on the industrial ICT ecosystems. Apart from the above, public administrations and research organizations are also using cloud. The usage of cloud is increasing in the domain of scientific computing and research works. We can execute various non-transient tasks on cloud. In other words we can mention that, it is possible to run virtual laboratories, virtual databases, and so on. These can be strongly coupled with the complex execution environments. In case of the conventional computing schemes just like batch systems or grid computing, the OS is usually imposed and all of the tools are needed to be managed properly outside the infrastructure.

All the resources in case of cloud computing are usually shared among various cloud clients with the help of virtualization theory. Virtualization technology is considered as the most important technology of data cloud centers which permits the dynamic sharing of physical resources. It also allows various applications to be executed in various platforms known as virtual machines. The quality of service can be enhanced with the help of this above mentioned virtualization theory.

By implementing the virtualization theory, maximum resource utilization and minimum power consumption can be obtained.

Information leakage and malicious alteration of sensitive data are considered as two important problems during the process of application deployment within a cloud. Security and performance are considered as the primary concern of every individual customer. Hence, there is necessity of secure and immediate data transfer and storage. We can mention here that, cloud is a combination of different interconnected computers and it has multiple united computing resources. With the growing popularity of cloud computing, customers can get high quality services all over the world. Clouds do provide computational and storage services on the basis of pay-per-use rule.

Presently, the most popular and widely acceptable internet-based model is no doubt cloud computing. It can provide access to shared pool of configurable assets on-request. These accesses can be provided and discharged with the providers' cooperation. Here we can mention here that, all applications can be executed on a virtual platform and each individual resource is distributed in between different virtual machines. Because of the resource parallelization, the utilization of CPU resources can be reduced to a great extent. As we all know if the tasks are not scheduled appropriately, then the performance can be degraded significantly. The process of scheduling plays important role in the domain of cloud computing. There is necessity of an algorithm in order to plan the task along with greatest evaluated gain. The computing ability of the distributed systems depends upon the costs of the resource utilized. Several distributed computing administrative tasks just like stockpiling and data transfer are quite simple to manage and these are helpful in order to reduce the costs. In case of clusters, there exists comparatively slow communication channel among processors. Therefore, the data communication process becomes more costly as compared to the supercomputers. There have been extensive amount of research works carried out in the field of scheduling in distributed computer systems. Along with the latest advancements of computer networks, the connecting links in between various computing entities have become faster. Here we can mention that, the latest applications require extended bandwidth, large storage and exchange of huge volumes of data. There are two important applications such as, multimedia and e-Science hose require huge volume of data. It is very much necessary to achieve better performance and quality of service.

Presently, large numbers of complex data streams are generated by different applications just like multimedia, social media, Internet of Things and social dispersed computing. In order to resolve the issues and support large scale data processing, different methodologies are adopted that supports the concepts of parallelism and big data. Most of these applications are implemented in cloud or multi-cloud environment. Cloud has virtually unlimited storage and computation resources which can be provided to any customer on demand. Mostly, the cloud customers want to decrease their expenses and delays through combination of privately owned systems with external public infrastructure. Again we can also state that, the cloud customers are more willing to enhance the process of resource utilization and throughput along with their monetary profits. Hence, in order to determine different cost-effective solutions and obtain the required service levels, scheduling of cloud resources is very vital. Cloud computing is the special domain which is more popular as compared to the distributed and grid computing. It involves the concepts of resource virtualization which is the most attractive feature of cloud. The term "computing" means the execution of different tasks on various virtual machines in order to carry out the process of efficient execution. With the growth of information and technology, the computational power of networks are also enhancing every day. The cloud behaves just like a repository of resources providing the customers with broader range of capabilities and computation facilities such as storage, processing, extraction and retrieval of the information. Apart from this, we can manage large numbers of heterogeneous tasks in order to generate better access. Basically, the pay-as-you-go or pay-as-you-use schemes are followed in cloud in order to access resources.

According to a recent survey, we can mention here that, we can get resource and task scalability, on time resource execution, dynamic provisioning, fault tolerance and interoperability of the resources from cloud. Apart from these, it has the responsibility to dynamically allocate cloudlets or tasks to the virtual machines. By implementing the load scheduling algorithm on cloud, dynamic allocation can be obtained easily and efficiently. It has the responsibility to achieve optimized throughput, reduced amount of execution and waiting time, reduced transfer time and decreased computational costs. The process of virtual machine scheduling and utilization is an emerging issue of distributed computing which can be categorized under the category of NP-Hard/NP-Complete problem. Cloud computing is considered as the most important technology of distributed environment that includes different data centers, servers, virtual machines, load balancers, and so on. All of these above mentioned entities are interlinked in a specific manner. Additionally, it can manage different other operations just like storage and extraction of documents, sharing of multimedia, lending the related resources on a pay-as-you-go model, etc. in future, additional research efforts can be performed in order to improve Infrastructure as a Service (IaaS) based cloud platform. Therefore, it is very much necessary to resolve the above mentioned issues. As all of these schemes are categorized under the category of NP-hard or NP-complete, we should give emphasis on the appropriate scheduling of virtual machines.Because of the flexibility and elasticity characteristics, the popularity of cloud computing is growing day by day. Large quantities of data are generated every day. Hence, it is very much difficult to handle and control those data efficiently. The healthcare domain is getting very much benefitted because of the advancement of cloud computing. This domain needs vast amount of resources. The healthcare industry is required to be automated in order to cope with the latest trends. Purchase and deploy high-end computer system for various jobs are very costly. All of these above mentioned issues can be resolved through the implementation of cloud technology.

Presently, cloud environments are used in order to deploy and execute various software applications more specifically FOSS-based applications. The above mentioned applications can include different cloud components in order to use resources. Horizontal and vertical elasticity is considered as the most attractive property of the SaaS level provided by cloud. Virtual elasticity has the responsibility to either increase or decrease the software resources abilities. On the other hand, horizontal elasticity has the responsibility to replicate or eliminate software instances. In case of globalized online services, large numbers of customers request are processed daily. The prime objective of application service provider is to process large numbers of requests in order to satisfy quality of service. Cloud computing is implemented in order to result decreased amount of IT expenses. Application service provider usually rent resources from various data centers in heterogeneous clouds. All of these data centers may be included within various cloud computing service providers. Basically, the nearest data center provides better quality of service. The cloud has the responsibility to offer resources on demand. It may require small amount of time to initiate a new virtual machine in order to satisfy the computing requirements. The existing computing resources are not capable to match the current demand. In case of highly dynamic resources, two major issues often arise, those are:- resource insufficiency and resource overload.

II. RELATED WORKS

H. Jiang, J. Yi, S. Chen and X. Zhu presented a multi-objective scheme in order to carry out the process of task scheduling and resource allocation in cloud-based disassembly [16]. There are certain manufacturers who do outsource the disassembly tasks to various professional organizations. Every individual organization has its own spec1ifications for the disassembly. Various kinds of disassembly facilities are integrated in order to execute different disassembly tasks. In this piece of research work, they presented a new cloud-based disassembly process in order to abstract the capability of the disassembly organization. After that, the disassembly resource is allocated to the tasks. Depending upon the above mentioned concepts, a new cloud-based disassembly system is presented. It is responsible for providing disassembly services based on the users' requirement. There are two major objectives of this system, those are:- decreasing the estimated total makespan and decreasing the estimated total expenses.

S. Jlassi, A. Mammar, I. Abbassi and Md. Graiet emphasized on efficient cloud resource allocation in case of different FOSS applications [17]. The cloud computing technology is used to construct demand free open source software (FOSS) applications. Because of the insufficient formal descriptions of previously developed traditional FOSS applications, the correctness of cloud resources management is not verified. The prime objective of this research paper is to introduce a formal definition in order to verify the correctness and consistency of cloud resource allocation. They have presented a new cloud resource allocation model with the help of event-B scheme. The verification process of correctness and consistency can be divided into two important phases, those are:-

1. The ProB model verifier has the responsibility to identify the most obvious errors and validate the event-B scheme.
2. After that, a proof activity is carried out in order to discharge the produced proof obligations. It has the responsibility to verify the correctness of the scheme.

G. Kaur, A. Bala and I. Chana proposed an intelligent regressive ensemble technique in order to predict resource usage in cloud computing [18].

A. A. Khan, M. Zakarya and R. Khan introduced energy-aware dynamic resource management scheme in case of elastic cloud data centres [19]. The placement procedure of on-demand applications on heterogeneous machines is considered as the major issue. There are numbers of different approaches those are only beneficial for reducing the energy consumption of a particular machine.These methods can't be implemented in order to obtain power optimization in High Performance Computing (HPC) systems. In this piece of research work, different resource management approaches are discussed. They have also suggested different heuristic techniques in order to optimise energy consumption and performance in elastic datacenters.

W. Kong, et.al, Ma introduced a new virtual machine resource scheduling technique for cloud computing-based on auction mechanism [20].

Y. C. Lee, et.al, on the resource-efficient workflow scheduling in clouds [21]. Because of the inefficient resource utilization during the execution of scientific workflows, there is need of an effective solution. In this research paper, they have identified the issue of resource-efficient workflow scheduling. They have introduced Maximum Effective Reduction (MER) scheme.

F. Li, et.al, developed a new two-level multi-task scheduling scheme in a cloud manufacturing environment [22]. In case of a cloud manufacturing environment, the issue of scheduling multiple heterogeneous tasks is very complicated. There exist various kinds of functional requirements during the tasks submission and there also exist the associated complications. Initially, it is subdivided into fine grained subtasks along with their precedence relationship. It is advisable to complete every individual task within the shortest time without any delay. On the contrary, in case of the cloud environment, every individual task is required to be completed as soon as possible. It is essential to schedule all subtasks from various heterogeneous tasks in order to enhance the advantages. In this research work, an advanced two-level multi-task scheduling scheme is presented. We can conclude here that, this scheme is more efficient as compared to the classical single-level scheme.

W. Lin, et.al, introduced an advanced heuristic task scheduling approach that depends upon sever power efficiency model [23]. The energy conservation in cloud datacenters has become the prime concern of various researchers. In this research paper, they have introduced a new power efficiency model for cloud servers. By implementing the above presented model, they used server power efficiency in order to assist the task scheduling process. They have presented a new heuristic task scheduling scheme in order to optimize energy conservation in cloud. This algorithm involves various multiple key factors just like task resource needs, server power efficiency model and performance degradation to decrease system energy consumption along with improved performance. This algorithm results very low time and space complexity along with enhanced global searching capability.

Y. Liu, W. Wei and R. Zhang introduced an effective differential evolution scheme for stochastic demand-oriented resource placement in heterogeneous clouds [24]. In case of geographically dispersed online services, users send request from any place. Depending upon the distributed cloud environment, the service provider is required to identify the optimal resource placement in order to enhance the revenue under constraints. Demand stochasticity and pricing heterogeneity are two vital factors those influence the problem's complexity to a great extent. In order to resolve the above mentioned issue, they have presented an effective differential evolution method for stochastic demand-oriented resource placement. In future, this approach can be modified and extended in order to achieve better performance.

Y. Lu, C. Lin, K. Lai, M. Tsai, Y. Wu, H. Chang and K. Huang focused on service deployment and scheduling in order to enhance the performance of composite cloud services [25]. The advancement of cloud computing have influenced the development and usage of software significantly. On the other hand, it also raised different research problems. In this piece of research work, they have discussed two important problems, those are:- service deployment and service request scheduling.They have proposed an advanced load-aware service deployment technique for dynamic workload. Again, they have introduced another service request scheduling scheme depending upon the concepts of tasks ranking mechanisms. The above presented technique is capable enough to enhance the overall execution performance.

S. H. H. Madni, et.al, introduced a new resource scheduling algorithm for Infrastructure as a Service (IaaS) in Cloud Computing [26]. The process of resource scheduling has the responsibility to assign the most appropriate task to the CPU, network, and storage. The prime objective of this algorithm is the extreme usage of resources. It is very much necessary to develop a proper scheduling mechanism for both cloud users and cloud service providers. In this research paper, they have analysed various resource scheduling techniques and classified them depending upon the problem. The parameters used during the evaluation process also play very important role. According to a recent survey, we can mention here that, most of the pre-existing approaches never include vital parameters those are very much required to enhance the performance. In the subsequent time, further research works can be performed in order to extend this technique.

L. Mao, et.al, proposed a new mulit-resource task scheduling technique for energy-performance trade-offs in green clouds [27]. Most of the previously developed task scheduling approaches emphasize on either energy conservation or improving performance. With the advancements of cloud computing, the users' requirements are becoming more complicated and diversified. This research work includes two important schemes, those are:- a time-aware scheme and an energy-aware scheme. Both of these schemes are developed to work on the heterogeneous environment. In the next phase, they integrated both of these schemes in order to build a new scheme and termed it as Energy-Performance Trade-Off Multi-Resource Cloud Task Scheduling scheme. Users are allowed to handle and control both energy and performance of cloud system through tuning the probability parameter α. This approach has the prime objective to decrease energy consumption and also decrease the completion time.

S. K. Panda, et.al, introduced an advanced pair-wise task scheduling approach in the cloud computing environment [28]. In cloud computing, task scheduling process plays very important role.All the scheduling algorithms have the basic objective to form an appropriate scheduling plan for all of the tasks. According to most of the researchers, task scheduling is considered as NP-complete problem and it has the primary goal to result minimum execution time. In this piece of research work, they have introduced a new pair-based task scheduling scheme. The above presented technique completely depends upon an optimization scheme known as Hungarian algorithm. It includes an unequal number of tasks and clouds. It makes pairs of tasks in order to take any scheduling decision.

V. Priya, et.al, presented a new resource scheduling scheme along with load balancing for cloud service provisioning [29]. In order to carry out the process of virtualized file sharing in cloud, the scheduling and load balancing processes are very important. Both of the above mentioned processes are required to be executed in an optimized manner. In order to achieve better load balancing and quality of service provisioning, scalable traffic management scheme is introduced. But, there exist numbers of different limitations in case of multidimensional resource allocation. Therefore, it is very much essential to develop an efficient resource scheduling in order to perform resource optimization. The major goal of this approach is to merge the process of resource scheduling along with load balancing. The use of virtual machines is increased with proper load balancing through dynamic selection of request with the help of Multidimensional Queuing Load Optimization scheme. Underutilization and overutilization of resources can be avoided through implementation of this approach.

P. Salza and F. Ferrucci emphasized on speeding up genetic algorithms in the cloud using software containers [30]. Because of scalability problems, genetic algorithms cannot be implemented in case of real world problems. Parallel computation is considered as the best option to resolve this issue. But in case of parallel computation, the overhead for communication is very high. They have introduced an optimized genetic algorithm in the form of a cloud-based application. The basic concepts of this approach depend upon master/ slave model. They devised conceptual workflow including the distribution phase, actual deployment and execution phase. In future, additional research works are needed to be carried out in order to modify and extend this above presented approach.

III. PROPOSED APPROACH

Improved PSO for optimal feature selection

Initializing particles with cloud services, number of iterations, velocity, number of particles etc.

Compute hybrid velocity and position for each particle in ‘d’ dimensions using the following equations

$$v(d+1,i) = \psi.[\omega(d,i).\,v(d,i).\theta_{chaos1}(pBest_i - X(d,i)) + \theta_{chaos2}(gBest_i - X(d,i))]$$

$$X(d+1,i) = X(d,i) + v(d+1,i)$$

ψ is the convergence factor computed as

$$\psi = \frac{2*(\theta_{c1}+\theta_2)}{|\,2-(\theta_{c1}+\theta_{c2})-\sqrt{(\theta_{c1}+\theta_{c2})^2-4(\theta_{c1}+\theta_{c2})}\,|}$$

where $\theta_{c1}, \theta_{c2} \in$ random numbers

1) $\delta_p -> \{\delta_{p_1}, \delta_{p_2}, \delta_{p_3} ... \delta_{p_m}\}$ represents the set of physical machines in cloud environment.

2) $\chi_C -> \{\chi_{C_1}, \chi_{C_2}, \chi_{C_3} ... \chi_{C_N}\}$ represents the set of data centres with resources in cloud environment.

$R_i = \{RAM, NBW, CPU, PROTO, ACCPOLI\}$

Where NBW: Network bandwidth, PROTO: PROTOCOL, ACCPOLI: Access policies.

3) $\phi_{VM} -> \{\phi_{VM_1}, \phi_{VM_2}, \phi_{VM_3} ... \phi_{VM_K}\}$ represents the set of cloud instances in cloud environment. Let $r_i \in R_i$ be the set of resources of all cloud instances.

$$r_i \in R_i \rightarrow \{r_i^{RAM}, r_i^{NBW}, r_i^{CPU}, r_i^{PROTO}, r_i^{ACCPOLI}\}$$

$$\begin{aligned} LIK^g_{a\rightarrow b} &= \log\left(P\left(S^g_{a,1},\ldots,S^g_{a,n_a} \mid L_a\right)\right) - \log\left(P\left(S^g_{a,1},\ldots,x^g_{a,n_a} \mid L_b\right)\right) \\ &= \log\left(\prod_{i=1}^{n_a} P\left(S^g_{a,i} \mid L_a\right)\right) - \log\left(\prod_{i=1}^{n_a} P\left(S^g_{a,i} \mid L_b\right)\right) \\ &= \sum_{i=1}^{n_a}\left(-\log\left(\varphi^g_a\right) - \frac{\left(S^g_{a,i}-\mu^g_a\right)^2}{\left(\varphi^g_a\right)} + S^g_{a,i}.\log\left(\varphi^g_b\right) + \frac{\left(S^g_{a,i}-\mu^g_b\right)^2}{2\left(\varphi^g_b\right)^2}\right) \end{aligned}$$

$$LoadOpt = -LIK^g_{a\rightarrow b}\,\frac{\partial F}{\partial w^{(2)}_{kj}} = LIK^g_{a\rightarrow b}\left\{\sum_{p=1}^{n}\left(\frac{d^p_k}{y^{(2),p}_k} - \frac{1-d^p_k}{1-y^{(2),p}_k}\right)\varphi^{(2)}\left(v^{(2),p}_k\right)\cdot y^{(1),p}_j\right\}$$

For each physical machine PM(δ_p), the Boolean bit vector $B_j = \{B_{j1}, B_{j2}, B_{j3} ..... B_{jr}\}$,

$$B_{jr} = 1 \text{ if } \phi_{VMj} \text{ assigns } \delta_{pr}.$$
$$B_{jr} = 0 \text{ otherwise}$$

Similarly the status of physical machine is denoted by { PB_m} where

$PB_m=1$ $(\exists\ \phi_{VM_i} \in \phi_{VM} / B_{mi=1})$

In this optimized model, inertia weight is computed as

$$\omega(d,i) = \omega_{max} - (I_{current} / I_{max}).(\omega_{max} - \omega_{min})$$
$$\omega_{max} : \max \text{inertia}$$
$$\omega_{min} : \min \ \text{inertia}$$
$$I_{max} : \max \ \text{iteration}$$

Step 3: Computing fitness value using ortho chaotic gauss randomization measure.

In this proposed PSO model, a random value between 0 to 1 is selected using the following equation as

$$R_i = \frac{1}{\sqrt{2\pi\sigma_X}} e^{-\frac{(X-\mu_X)^2}{\sigma_X{}^2}}$$

$K = 1, 2...\text{iterations}$

Proposed Objective Functions for Non-linear constraint programming:

The objective function for the proposed resource optimization in cloud computing environment is given by

1) $$\text{Min} \sum_{i=1}^{N} \log\{PB_i\} \text{ and Max } \sum_{j=1}^{M} \{B_j\}$$
$$\text{with } PB_i = 1((\exists\ \phi_{VM_i} \in \phi_{VM} / B_{ij=1})$$
$$PB_i = 0; \text{otherwise}$$
---(1)

2) $$\text{Max} \sum_{j=1}^{M} \{B_j\}$$
$$B_{jr} = 1 \text{ if } \phi_{VMj} \text{ assigns } \delta_{pr}.$$
$$B_{jr} = 0 \text{ otherwise}$$
-----(2)

S.t

$$\sum_{p=1}^{|V|} r_p^{RAM}.PB_{ip} \le C_i^{RAM}; 1 \le i \le N$$
$$\sum_{p=1}^{|V|} r_p^{NBW}.PB_{ip} \le C_i^{NBW}; 1 \le i \le N$$
$$\sum_{p=1}^{|V|} r_p^{CPU}.PB_{ip} \le C_i^{CPU}; 1 \le i \le N$$
$$\sum_{p=1}^{|V|} r_p^{PROTO}.PB_{ip} \le C_i^{PROTO}; 1 \le i \le N$$
$$\sum_{p=1}^{|V|} r_p^{ACCPOLI}.PB_{ip} \le C_i^{ACCPOLI}; 1 \le i \le N$$

Energy Constraint:

Energy consumption of computing resources such as job computation , server storage, server capacities can be computed using the power model. The linear relationship among the resource utilisation and power consumption is given as:

$$PU\left(CPU_i\right) = \alpha P_{max}(CPU_i) + (1-\alpha).P_{max}(CPU_i).PU\left(CPU_i\right)$$

$PU\left(CPU_i\right)$ is the power utilisation of cloud instance i.

$P_{max}(CPU_i)$ is the maximum power utilisation of ith instance , when the cloud server is fully allocated.

α is the fraction of scaling parameter to the idle server (0-1).

In multi-core cloud environment the total utilization of all cloud instances should be minimized using the constraint programming as

$$SPU\left(CPU_i\right) = \sum_{i=1}^{N} PU\left(CPU_i\right)$$

$$\min SPU\left(CPU_i\right) = \min \sum_{i=1}^{N} PU\left(CPU_i\right)$$

Total power utilisation of al lthe cloud instances in the available data centre is given by

$$TP = \alpha P_{max}(CPU_i) + (1-\alpha)SPU\left(CPU_i\right)$$

The energy consumption of all the cloud server over a time period T is given by

$$TP_i = \int_0^T TP(t)\log(TP(t))dt \qquad \text{----(3)}$$

Proposed feature selection fitness measure is given as

$$Fitness_i = w_1.TP_i + w_2.(1 - \frac{\sum_{i=1}^{|F|} f_i}{N_f})$$

where $w_1, w_2 \in R_i$

f_i is the flag value 1 or 0 . '1' represents selected service , '0' non-selected resource.

N_f represents number of services.

Step 4: For each particle compute its fitness value and compute classification accuracy in the previous step.

Step 5: Update particle velocity, position, global best and particle best according to the fitness value conditions.

Step 6: This process is continuous until max iteration is reached. Otherwise go to step 2.

The main objective of the proposed resource optimization model is to minimize the number of physical machine required to host all the instances.

Algorithm Steps:

1. Connect to cloud environment using credentials with available data centre zones.
2. Initialization of ‘k1’ number of available data centre zones DC[].
3. Initialization of ‘k2’ number of physical machines PM[].
4. Initialization of ‘k3’ number of virtual instances VI[].
5. For each user request of instance VI[i]
6. Do
7. Search PM in the available data centres DC[].
8. Search for instance capacity and its properties in the physical machine PM[].
9. Check the optimization functions for the data centres, Physical machine, virtualmachines and energy computation using (1),(2),(3).
10. Estimating the best servers using the proposed probability estimation formula. The minimum and maximum bound limits are used to decide the workload usage of each instance in the virtual machine as:

$$\text{Lower bound limite :}\mu_{VI[]} - \lambda\sigma_{VI[]}$$

$$\text{Upper bound limite :}\mu_{VI[]} + \lambda\sigma_{VI[]}$$

$$\text{Bounded limit :}\ \mu_{VI[]}$$

$$\lambda = \frac{1}{\sigma_{VI[]}.\sqrt{2.\pi}} e^{-\frac{(VI[i]-\mu_{VI[]})^2}{2.\sigma^2_{v[]}}}$$

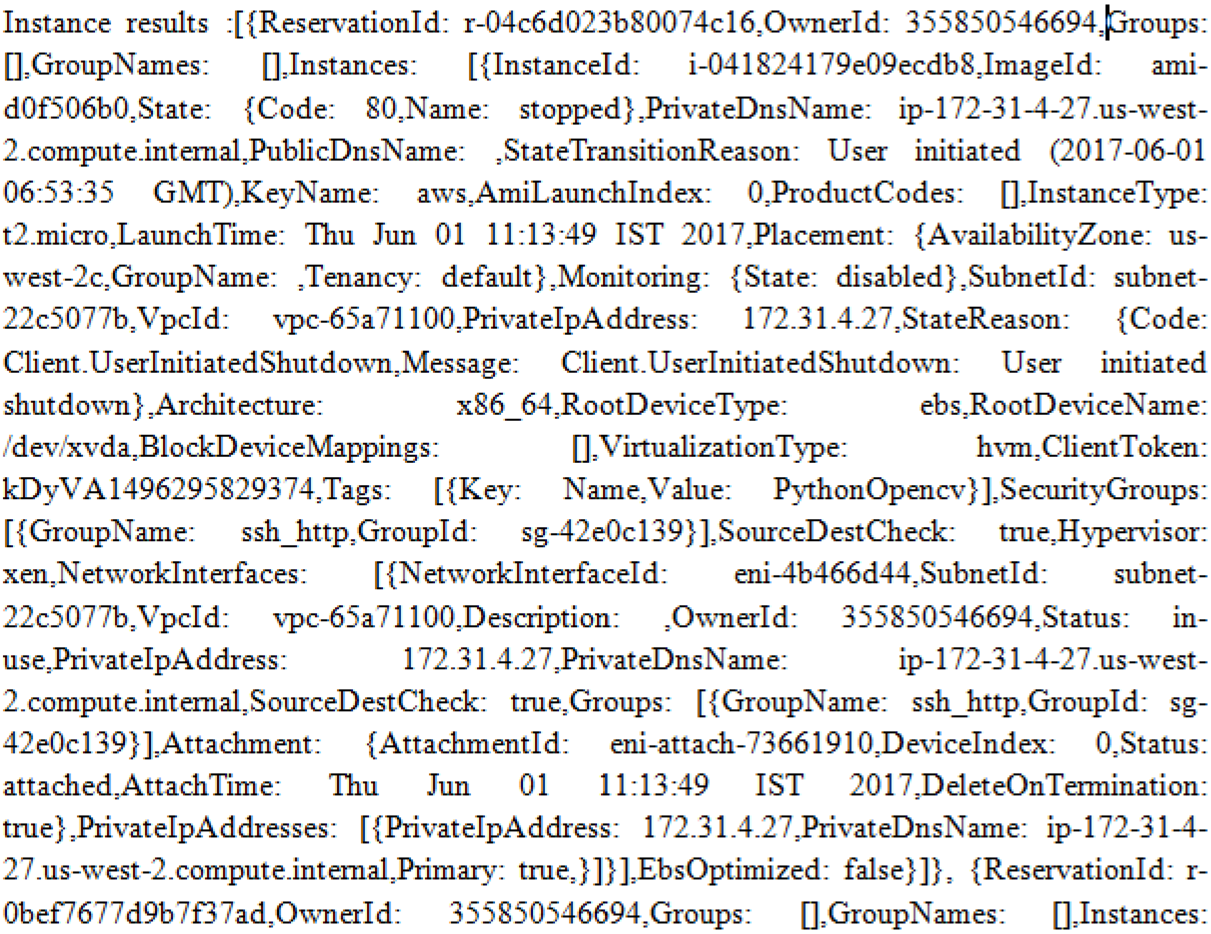
$$\text{minimize} \quad \mathbf{w}^T\mathbf{w} + \text{LoadOpt}\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad y_i\left(\mathbf{w}^T\mathbf{x}_i + \text{LIK}^{g}_{a\to b}\right) \geq 1 \quad i = 1,\ldots,n$$

IV. EXPERIMENTAL RESULTS

For experimental results, homogeneous and heterogeneous virtual machines have been used that consist of five instances with specified number of resources and data. To compare the performance of the existing models with the proposed model, three metrics have been used to evaluate the load balancing, energy consumption and runtime of the virtual instances and available resources. For virtual machine, kernel based VM has been installed in each server node in cloud environment. Different operating systems such as Red hat linux, Centos, Windows etc are used to evaluate the performance of each virtual machine in the cloud environment. For experimental evaluation, Amazon aws cloud environment is used to test the optimial resource allocation and to test the efficiency of the proposed model to the existing models. All experimental results are performed using the Java programming environment with real-time amazon aws third party libraries.

The initialization of the cloud instances and its resources are summarized below:



```
Instance results :[{ReservationId: r-04c6d023b80074c16,OwnerId: 355850546694,Groups:
[],GroupNames: [],Instances: [{InstanceId: i-041824179e09ecdb8,ImageId: ami-
d0f506b0,State: {Code: 80,Name: stopped},PrivateDnsName: ip-172-31-4-27.us-west-
2.compute.internal,PublicDnsName: ,StateTransitionReason: User initiated (2017-06-01
06:53:35 GMT),KeyName: aws,AmiLaunchIndex: 0,ProductCodes: [],InstanceType:
t2.micro,LaunchTime: Thu Jun 01 11:13:49 IST 2017,Placement: {AvailabilityZone: us-
west-2c,GroupName: ,Tenancy: default},Monitoring: {State: disabled},SubnetId: subnet-
22c5077b,VpcId: vpc-65a71100,PrivateIpAddress: 172.31.4.27,StateReason: {Code:
Client.UserInitiatedShutdown,Message: Client.UserInitiatedShutdown: User initiated
shutdown},Architecture: x86_64,RootDeviceType: ebs,RootDeviceName:
/dev/xvda,BlockDeviceMappings: [],VirtualizationType: hvm,ClientToken:
kDyVA1496295829374,Tags: [{Key: Name,Value: PythonOpencv}],SecurityGroups:
[{GroupName: ssh_http,GroupId: sg-42e0c139}],SourceDestCheck: true,Hypervisor:
xen,NetworkInterfaces: [{NetworkInterfaceId: eni-4b466d44,SubnetId: subnet-
22c5077b,VpcId: vpc-65a71100,Description: ,OwnerId: 355850546694,Status: in-
use,PrivateIpAddress: 172.31.4.27,PrivateDnsName: ip-172-31-4-27.us-west-
2.compute.internal,SourceDestCheck: true,Groups: [{GroupName: ssh_http,GroupId: sg-
42e0c139}],Attachment: {AttachmentId: eni-attach-73661910,DeviceIndex: 0,Status:
attached,AttachTime: Thu Jun 01 11:13:49 IST 2017,DeleteOnTermination:
true},PrivateIpAddresses: [{PrivateIpAddress: 172.31.4.27,PrivateDnsName: ip-172-31-4-
27.us-west-2.compute.internal,Primary: true,}]}],EbsOptimized: false}]}, {ReservationId: r-
0bef7677d9b7f37ad,OwnerId: 355850546694,Groups: [],GroupNames: [],Instances:
```

```
[],Instances: [{InstanceId: i-0b1dc4321d02370d5,ImageId: ami-58998521,State: {Code:
80,Name: stopped},PrivateDnsName: ip-172-31-45-202.us-west-
2.compute.internal,PublicDnsName: ,StateTransitionReason: User initiated (2018-01-15
11:51:46 GMT),KeyName: gskpair,AmiLaunchIndex: 0,ProductCodes: [],InstanceType:
t2.micro,LaunchTime: Mon Jan 15 17:19:32 IST 2018,Placement: {AvailabilityZone: us-
west-2b,GroupName: ,Tenancy: default},Monitoring: {State: disabled},SubnetId: subnet-
63e36d06,VpcId: vpc-65a71100,PrivateIpAddress: 172.31.45.202,StateReason: {Code:
Client.UserInitiatedShutdown,Message: Client.UserInitiatedShutdown: User initiated
shutdown},Architecture: x86_64,RootDeviceType: ebs,RootDeviceName:
/dev/sda1,BlockDeviceMappings: [{DeviceName: /dev/sda1,Ebs: {VolumeId: vol-
0016c75b7283d6c37,Status: attached,AttachTime: Mon Nov 20 20:11:49 IST
2017,DeleteOnTermination: true}}],VirtualizationType: hvm,ClientToken: ,Tags: [{Key:
Name,Value: GSKSPARKJAVA}],SecurityGroups: [{GroupName: launch-wizard-
2,GroupId: sg-d48560a8}],SourceDestCheck: true,Hypervisor: xen,NetworkInterfaces:
[{NetworkInterfaceId: eni-bfb4c79f,SubnetId: subnet-63e36d06,VpcId: vpc-
65a71100,Description: ,OwnerId: 355850546694,Status: in-use,PrivateIpAddress:
172.31.45.202,PrivateDnsName: ip-172-31-45-202.us-west-
2.compute.internal,SourceDestCheck: true,Groups: [{GroupName: launch-wizard-2,GroupId:
sg-d48560a8}],Attachment: {AttachmentId: eni-attach-c533b525,DeviceIndex: 0,Status:
attached,AttachTime: Mon Nov 20 20:11:49 IST 2017,DeleteOnTermination:

true},PrivateIpAddresses: [{PrivateIpAddress: 172.31.45.202,PrivateDnsName: ip-172-31-
45-202.us-west-2.compute.internal,Primary: true,}]}],EbsOptimized: false}]},
{ReservationId: r-0eeb2df62550d7c0f,OwnerId: 355850546694,Groups: [],GroupNames:
[],Instances: [{InstanceId: i-047f013f8b88ef80d,ImageId: ami-82ccade2,State: {Code:
80,Name: stopped},PrivateDnsName: ip-172-31-33-232.us-west-
2.compute.internal,PublicDnsName: ,StateTransitionReason: User initiated (2017-06-02
05:33:39 GMT),KeyName: aws,AmiLaunchIndex: 0,ProductCodes: [],InstanceType:
t2.micro,LaunchTime: Fri Jun 02 11:03:11 IST 2017,Placement: {AvailabilityZone: us-west-
2b,GroupName: ,Tenancy: default},Monitoring: {State: disabled},SubnetId: subnet-
63e36d06,VpcId: vpc-65a71100,PrivateIpAddress: 172.31.33.232,StateReason: {Code:
Client.UserInitiatedShutdown,Message: Client.UserInitiatedShutdown: User initiated
shutdown},Architecture: x86_64,RootDeviceType: ebs,RootDeviceName:
/dev/sda1,BlockDeviceMappings: [],VirtualizationType: hvm,ClientToken:
poZBQ1496231627480,Tags: [{Key: Name,Value: RStudioWebGSK}],SecurityGroups:
[{GroupName: ssh_http,GroupId: sg-42e0c139}],SourceDestCheck: true,Hypervisor:
xen,NetworkInterfaces: [{NetworkInterfaceId: eni-0787452d,SubnetId: subnet-
63e36d06,VpcId: vpc-65a71100,Description: ,OwnerId: 355850546694,Status: in-
use,PrivateIpAddress: 172.31.33.232,PrivateDnsName: ip-172-31-33-232.us-west-
2.compute.internal,SourceDestCheck: true,Groups: [{GroupName: ssh_http,GroupId: sg-
42e0c139}],Attachment: {AttachmentId: eni-attach-e2215c0b,DeviceIndex: 0,Status:
attached,AttachTime: Wed May 31 17:23:48 IST 2017,DeleteOnTermination:
true},PrivateIpAddresses: [{PrivateIpAddress: 172.31.33.232,PrivateDnsName: ip-172-31-
33-232.us-west-2.compute.internal,Primary: true,}]}],EbsOptimized: false}]}]You have 4
Amazon EC2 instance(s) running.
```

The complexity of proposed model to the existing models depends on the number of physical machines and virtual machines. In the experiments, different number of physical machines and virtual machines are used to measure the improved of the proposed model to the existing models. The maximum and minimum bound limits of the physical machines and virtual machines computed in the proposed model are listed in table 1 and table 2.

CPU Utilization (Percent)
Disk Reads (Bytes)
Disk Read Operations (Operations)
Disk Writes (Bytes)
Disk Write Operations (Operations)
Network In (Bytes)
Network Out (Bytes)
Network Packets In (Count)
Network Packets Out (Count)
Status Check Failed (Any) (Count)
Status Check Failed (Instance) (Count)
Status Check Failed (System) (Count)
1
0.75
0.5
0.25
0
5/3
08:00
5/3
08:30

Figure 1: Computational analysis of proposed model based on different datasets

Bounds	CPU(Hz)	RAM(MB)	BANDWIDTH(Kbps)
Lower bound	1500	1000	1500
Upper bound	9500	9000	9000

Table 1: Physical machine bound limits

Bounds	CPU(Hz)	RAM(MB)	BANDWIDTH(Kbps)
Lower bound	350	400	200
Upper bound	3500	3500	800

Table 2: Virtual machine bound limits

Table 3: **Comparative analysis of resource allocation and runtime of the proposed model to the existing models.**

Model	**Avg Allocated Resources**	**Runtime(ms)**
Papagianni et al, 2013	15	9743
Buyya et.al, 2016	12	8475
Jenning et.al, 2015	13	9164
Poola et.al,	16	7935
Multiobjective	9	7395
Loadbalancing model	9	7143
Proposed Model	8	6983

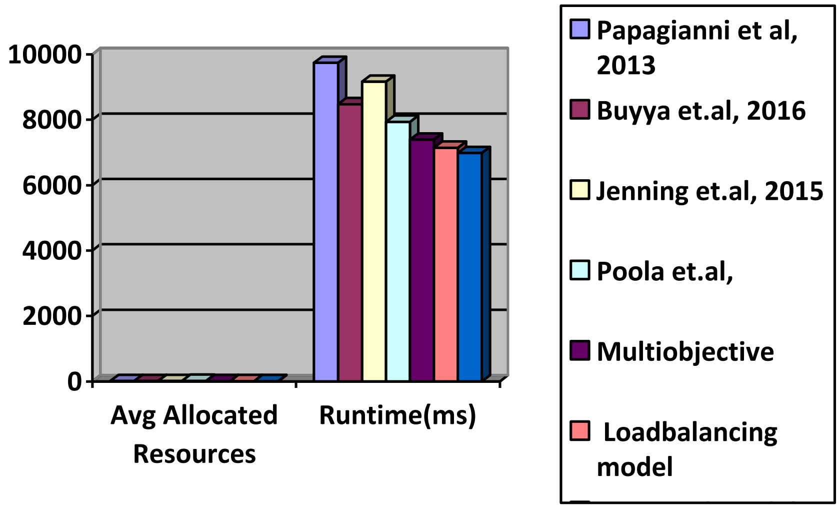


Figure 2 : Comparative analysis of runtime of the proposed model to the existing models.

V. CONCLUSION

In this paper, different load balancing functions are integrated by using cloud optimization functions. These models are designed and implemented to test the resource allocation using the available physical machines and virtual instances. Load balance can improve quality of service (QoS) measurements, including response time, cost, performance and use of resources. Since, most of the cloud service providers have their own type, type and price policies for computing resources, including other service features. Since, computational time and memory of the existing cloud scheduling models are not efficient in realtime cloud environment .The load balance between cloud resources ensures an efficient utilization of the physical infrastructure while minimizing runtime However, the main problem to the cloud service provider's is optimizing cloud service parameters such as reliability, flexibility, time limits and the task refusal rate. In order to overcome the cloud resource allocation and scheduling problems a novel cloud resource allocation and load balancer framework is used to improve the cloud allocation issues. Experimental results proved that the present load-balancing model has better performance than the traditional load balancing approaches on various cloud resources.

VI. REFERENCES

[1] H. Jiang, J. Yi, S. Chen and X. Zhu, A multi-objective algorithm for task scheduling and resourceallocation in cloud-based disassembly, Journal of Manufacturing Systems 41 (2016) 239–255

[2] S. Jlassi, A. Mammar, I. Abbassi and Md. Graiet, Towards correct cloud resource allocation in FOSS applications, future generation computer systems.

[3] G. Kaur, A. Bala and I. Chana , An intelligent regressive ensemble approach for predicting resource usage in cloud computing, journal of parallel and distributed computing.

[4] A. A. Khan, M. Zakarya and R. Khan, Energy-aware Dynamic Resource Management in Elastic Cloud Datacenters, Simulation Modelling Practice and Theory.

[5] W. Kong, Y. Lei and J. Ma, Virtual machine resource scheduling algorithm for cloud computingbased on auction mechanism.

[6] Y. C. Lee, H. Han, A. Y. Zomaya and M. Yousif, Resource-efficient workflow scheduling in clouds, knowledge based systems

[7] F. Li, T.W. Liao and L. Zhang , Two-level multi-task scheduling in a cloud manufacturing environment, Robotics and Computer Integrated Manufacturing 56 (2019) 127–139.

[8] W. Lin, W. Wang, W. Wu, X. Pang, B. Liu and Y. Zhang, A heuristic task scheduling algorithm based on server powerefficiency model in cloud environments, Sustainable Computing: Informatics and Systems.

[9] Y. Liu, W. Wei and R. Zhang, DESRP: An efficient differential evolution algorithm for stochastic demand-oriented resource placement in heterogeneous clouds, future generation computer systems.

[10] Y. Lu, C. Lin, K. Lai, M. Tsai, Y. Wu, H. Chang and K. Huang, Service deployment and scheduling for improving performance of composite cloud services, Computers and Electrical Engineering.

[11] S. H. H. Madni, M. S. A. Latiff, Y. Coulibaly and S. M. Abdulhamid, Resource Scheduling for Infrastructure as a Service (IaaS) in Cloud Computing: Challenges and Opportunities, Journal of Network and Computer Applications.

[12] L. Mao, Y. Li, G. Peng, X. Xu and W. Lin, A Multi-Resource Task Scheduling Algorithm for Energy-Performance Trade-offs in Green Clouds

[13] S. K. Panda, S. S. Nanda and S. K. Bhoi, A Pair-Based Task Scheduling Algorithm for Cloud Computing Environment, Journal of King Saud University - Computer and Information Sciences.

[14] V. Priya, C. S. Kumar and R. Kannan, Resource scheduling algorithm with load balancing for cloud service Provisioning, applied soft computing.

[15] P. Salza and F. Ferrucci, Speed up genetic algorithms in the cloud using software containers, future generation computer systems.