Abstract - Cloud Computing is a new technology which uses virtual machine instead of physical machine to host, store and network the different components. Cloud Computing is a model that points at streamlining the on-demand provisioning of software, hardware & data as services & providing end-users with flexible & scalable services accessible through the Internet. Load Balancing is one of the major issues in Cloud Computing. There are a number of soft computing techniques available to optimize the load. In this paper, some of them are discussed.

Index Terms—Cloud Computing, Load Balancing, Meta-heuristic.

I. INTRODUCTION

Distributed network computing environments have become a cost effective and popular choice to achieve high performance and to solve large scale computation problems. Unlike past supercomputers a cloud or cluster or grid system can be used as multipurpose computing platform to run diverse high performance parallel applications. Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a utility (like the electricity grid) over a network (typically the Internet). The term Cloud refers to a Network or Internet. Cloud can provide services over network, i.e., on public networks or on private networks, i.e., WAN, LAN or VPN. Applications such as e-mail, web conferencing, customer relationship management (CRM), all run in cloud.

A 'cloud' is an elastic execution environment of resources involving multiple stakeholders and providing a metered service at multiple granularities for a specified level of quality (of service). Cloud Computing refers to manipulating, configuring, and accessing the applications online. It offers online data storage, infrastructure and application. Cloud Computing is high utility software having the ability to change the IT software industry and making the software even more attractive. It has also changed the way IT companies used to buy and design hardware. The elasticity of resources without paying a premium for large scale is unprecedented in the history of IT industry. The increase in web traffic and different services are increasing day by day making load balancing a big research topic.

A. Characteristics of Cloud Computing

A Cloud is a form of parallel and distributed system possessing a group of inter-connected and virtualized computers that are dynamically scheduled and highlighted as one or more unified computing resources based on service-level agreements established through conciliation between the service provider and consumers [1]. There are various characteristics of Cloud Computing and includes on-demand self service, broad network access, resource pooling, rapid elasticity and measured service.

- On-demand self service means that customers (usually organizations) can request and manage their own computing resources.
- Broad network access allows services to be offered over the Internet or private networks.
- Pooled resources means that customers draw from a pool of computing resources, usually in remote data centres.
- Services can be scaled larger or smaller; and use of a service is measured and customers are billed accordingly.

B. Service Models of Cloud Computing

Cloud computing providers offer their services according to several fundamental models [1], which are shown below:

- **Infrastructure as a service (IaaS):** In this model, the cloud user patches and maintains the operating...
systems and the application software. Cloud providers typically bill IaaS services on a utility computing basis: cost reflects the amount of resources allocated and consumed.

b) **Platform as a service (PaaS):** In the PaaS models, cloud providers deliver a computing platform, typically including operating system, programming language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without the cost and complexity of buying and managing the underlying hardware and software layers.

c) **Software as a service (SaaS):** In this, users are provided access to application software and databases. Cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Cloud users do not manage the cloud infrastructure and platform where the application runs. SaaS is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis. SaaS providers generally price applications using a subscription fee.

d) Other cloud services include the following:
   1. **DaaS (Database as a Service):** Database systems provide a user friendly interface for accessing and managing data. This type of service is very useful like many financial, business, and Internet-based applications [14].
   2. **NaaS (Network as a Service):** With NaaS, providers offer customers a virtualized network [15].
   3. **IPMaaS (Identity and Policy Management as a Service):** With this service, providers deliver identity and policy management to customers [16].

**C. Deployment Models**

Deployment Models for cloud environment are given as:

a) **Private Cloud:** Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party, and hosted either internally or externally. Undertaking a private cloud project requires a significant level and degree of engagement to virtualize the business environment, and requires the organization to reevaluate decisions about existing resources. When done right, it can improve business, but every step in the project raises security issues that must be addressed to prevent serious vulnerabilities.

b) **Public Cloud:** A cloud is called a "public cloud" when the services are rendered over a network that is open for public use. Public cloud services may be free or offered on a pay-per-usage model. Public cloud service providers like Amazon AWS, Microsoft and Google own and operate the infrastructure at their data center and offer access only via Internet.

c) **Hybrid Cloud:** Hybrid cloud is a composition of two or more clouds (private, community or public) that remain distinct entities but are bound together, offering the benefits of multiple deployment models. Hybrid cloud can also mean the ability to connect collocation, managed and/or dedicated services with cloud resources.

**D. Benefits of Cloud Computing**

Cloud computing offers lots of advantages:

- **Cost:** As in the clouds the user do not own the resources, user just pays as per the usage in terms of time, storage and services. It reduces the cost of owning the infrastructure.
- **Performance:** The performance is improved because the cloud is not a single computer but a large network of powerful computers resulting in high processing power.
- **Freedom from up gradation and maintenance:** The cloud infrastructure is maintained and upgraded by the cloud service provider.
- **Scalability:** The user is can request to increase the resources if the area of application grows or new functionality is added. On the other hand if requirement shrinks the user can request to reduce the resources as well.
- **Speedy Implementation:** Time of Implementation of cloud for an application may be in days or sometimes in hours. You just need a valid credit card and need to fulfill some online registration formalities.
- **Green:** The cloud computing is a green technology since it enable resource sharing among users thus not requiring large data centers that consumes a lot of power.
- **Mobility:** We don’t need to carry our personal computer, because we can access our documents anytime anywhere.
- **Maximized Storage Capacity:** In Cloud computing we have extreme resources for storing data because our storage consists of many bases in the Cloud. Another thing about storing data in the Cloud is that, because of our data in the Cloud can automatically duplicated, they will be more safety.

Another thing about storing data in the Cloud is that, because of our data in the Cloud can automatically get duplicated, there will be more safety.

**II. LOAD BALANCING**

Cloud Computing is the most recent emerging paradigm promising to turn the vision of “computing utilities” into reality, it provides a flexible and easy way to store and retrieve huge data without worrying about the hardware needed.

As the number of users on cloud increases, the existing resources decreases automatically which leads to the problem of delay between the users and the cloud service providers.

Thus, the load balancing comes into the picture. The traffic over the network must be dealt smartly such that the situation
in which some nodes are overloaded and some other are under loaded should never arise. To overcome this situation, many load balancing algorithms are proposed by researchers, with their own pros and cons. In this paper, an overall review of the current load balancing algorithms in the Cloud Computing environment is presented.

Load balancing [4] is a generic term used for distributing a larger processing load to smaller processing nodes for enhancing the overall performance of system. Load Balancing is a process of reassigning the total load to the individual nodes of the collective system to make more resource utilization effective and to improve the response time of the job, simultaneously removing a condition in which some of the nodes are overloaded while some others are under loaded.

### TABLE I: METRICS IN LB TECHNIQUES IN CLOUD COMPUTING

<table>
<thead>
<tr>
<th>LOAD BALANCING METRICS IN CLOUD COMPUTING</th>
<th>ILLUSTRATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>Illustration</td>
</tr>
<tr>
<td>Throughput</td>
<td>It is used to calculate the number of tasks whose execution has been completed. It should be high to improve the performance of the system.</td>
</tr>
<tr>
<td>Overhead</td>
<td>It determines the amount of overhead involved while implementing a load balancing algorithm. It is composed of overhead due to movement of tasks, inter-processor and inter-process communication. This should be minimized so that a load balancing technique can work efficiently.</td>
</tr>
<tr>
<td>Fault Tolerance</td>
<td>It is the time taken to migrate the jobs or resources from one node to other. It should be minimized in order to enhance the performance of the system</td>
</tr>
<tr>
<td>Response Time</td>
<td>It is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.</td>
</tr>
<tr>
<td>Resource Utilization</td>
<td>It is used to check the utilization of resources. It should be optimized for an efficient load balancing.</td>
</tr>
<tr>
<td>Scalability</td>
<td>It is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.</td>
</tr>
<tr>
<td>Performance</td>
<td>It is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays</td>
</tr>
</tbody>
</table>

A load balancing protocol is dynamic in nature doesn’t contemplate the previous state or behavior of the system, that is, it depends on the current behavior of the system. It is common these days in redundant high-availability computer systems that incoming network traffic is distributed on network level by deploying one of the frequently used network load balancing algorithms like:- random-allocation, round-robin allocation, weighted round-robin allocation, etc. These algorithms use solely network parameters of incoming traffic to create selections wherever to forward traffic, with none data from different elements of database system, like current load of application or info servers. Since these days it is extremely common to possess internet servers acting as application servers, it is usual that load balancers use session-switching technique, which suggests that once a user opens website on one server, it will stay on it server whereas the session lasts.

An ideal load balancing algorithm should avoid overloading or under loading of any specific node. But, in case of a cloud computing environment the selection of load balancing algorithm is not easy because it involves additional constraints like security, reliability, throughput etc. So, the main goal of a load balancing algorithm in a cloud computing environment is to improve the response time of job by distributing the total load of system. The algorithm must also ensure that it is not overloading any specific node.

Load Balancing is done with the help of load balancers where each incoming request is redirected and is transparent to the client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server [5]. To make the final determination, the load balancer retrieves information about the candidate server’s health and current workload in order to verify its ability to respond to the request. Load Balancing helps in [2]:

- Improving the performance substantially.
- Having a Reverse up plan in case the system fails even partially.
- Maintenance of system stability.
- Accommodation of future modification.
- Efficient load distribution.
- Cost effectiveness.

Depending on who initiated the process, load balancing algorithms can be of five categories:

- **Sender Initiated:** If the load balancing algorithm is initialized by the sender
- **Receiver Initiated:** If the load balancing algorithm is initiated by the receiver
- **Symmetric:** It is the combination of both sender initiated and receiver initiated
- **Static:** It doesn’t depend on the current state of the system. Prior knowledge of the system is needed.
- **Dynamic:** Decisions on load balancing are based on current state of the system. No prior knowledge is needed. So it is better than static approach.

### A. Policies of load balancing algorithm

There are many policies are used in load balancing algorithms: [14] [15]

- Information policy: It defined that what information is required and how this information is collected.
This is also defined that when this information is collected
- Triggering policy: This policy defined that time period when the load balancing operation is starting to manage the load.
- Resource type policy: This policy defined the all types of resources which are available during the load balancing.
- Location policy: This uses all the results of the resource type policy. It is used to find a partner for a server or receiver.
- Selection policy: This policy is used to find out the task which transfers from overloaded node to free node.

B. Major goals of load balancing algorithms
1. Cost effectiveness: Load balancing help in provide better system performance at lower cost.
2. Scalability and flexibility: The system for which load balancing algorithms are implemented may be change in size after some time. So the algorithm must handle these types' situations. So algorithm must be flexible and scalable.
3. Priority: Prioritization of the resources or jobs needs to be done. So higher priority jobs get better chance to execute.

III. META-HEURISTIC BASED LOAD BALANCING SCHEMES
A. Genetic Algorithm:
A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a metaheuristic) is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover [20].

B. Simulated Annealing:
Simulated annealing (SA) is a generic probabilistic metaheuristic for the global optimization problem of locating a good approximation to the global optimum of a given function in a large search space. It is often used when the search space is discrete (e.g., all tours that visit a given set of cities). For certain problems, simulated annealing may be more efficient than exhaustive enumeration — provided that the goal is merely to find an acceptably good solution in a fixed amount of time, rather than the best possible solution [21].

C. Ant Colony Optimization:
Ant Colony Optimization (ACO) is the result of research on computational intelligence approaches to combinatorial optimization originally conducted by Dr. Marco Dorigo, in collaboration with Alberto Colorini and Vittorio Maniezzo. The fundamental approach underlying ACO is an iterative process in which a population of simple agents repeatedly construct candidate solutions; this construction process is probabilistically guided by heuristic information on the given problem instance as well as by a shared memory containing experience gathered by the ants in previous iteration [22].

D. Artificial Bee Colony Optimization:
The artificial bee colony algorithm (ABC) is an optimization algorithm based on the intelligent foraging behavior of honey bee swarm, proposed by Karaboga in 2005. The number of employed bees in the colony is equal to the number of food sources around the hive. Employed bees go to their food source and come back to hive and dance on this area. The employed bee whose food source has been abandoned becomes a scout and starts to search for finding a new food source. Onlookers watch the dances of employed bees and choose food sources depending on dances [23].

E. Particle Swarm Optimization:
Particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position but, is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions [24].

IV. LITERATURE REVIEW
In [3], authors proposed a resource allocation mechanism with pre-emptable task execution which increases the utilization of clouds. They proposed an adaptive resource allocation algorithm for cloud system with pre-emptable tasks but their approach did not pertain to cost optimization and time optimization. In [4], authors focused on the various load balancing algorithms and their applicability in cloud computing environment. They first categorized the algorithms as static and dynamic. Then the various algorithms were analyzed which could be applied in static environments. After that they described the various dynamic load balancing algorithms. For solving any particular problem some special conditions needed to be applied. Some additional algorithms were discussed which helped in solving some sub-problems in load balancing which were applicable to cloud computing.

In [5], authors proposed a new enhanced and efficient scheduling algorithm and then implemented in cloud computing environment using CloudSim toolkit, in java language. By visualizing the cited parameters in graphs and tables they easily identified that the overall response time and data centre processing time was improved as well as cost was reduced in comparison to the existing scheduling parameters.

In [6], authors described a survey on load balancing schemes in cloud environments. There were various load balancing techniques used in these papers and their corresponding advantages, disadvantages and performance metrics are studied in detail.

In [7], authors focused on consequences of inefficient load balancing might lead to detriments of an organization business performance on cloud environment. Hence they illustrated...
various aspects pertaining to domain of cloud computing, its evolution, its generic issues, and particularly to issues related to load balancing. Various techniques analyzed and the findings were illustrated. Authors discussed only two divisible load scheduling algorithms that could be applied to clouds, but there were still other approaches that could be applied to balance the load in clouds. The performance of the given algorithms could also be increased by varying different parameters in [18]. Authors proposed a system that was composed of several replicated machines called “Virtual redundant machine” that can unload the load once the data centers get overloaded. This fulfilled their purpose of attaining Virtual machine migration which was one of the major challenges for load balancing mechanism. Better Process Management, better Resource allocation and better response time to end users could be achieved, in [13].

In [17], authors discussed the concept of “cloud” computing, the issues it tries to address, related research topics, and a “cloud” implementation based on VCL technology. They ran on cloud computing platform, some databases are SQL-based while some are non-SQL and discussed a service-oriented architecture, reduced information technology overhead for the end-user, great flexibility, reduced total cost of ownership, on-demand services and many other things. In [19] introduced the concept, history, pros and cons of cloud computing as well as the value chain and standardization effort.

In [20], Evolutionary algorithms, their types, techniques and relevance to current trends and operation of most Search Engines, i.e. through optimization had been discussed. A novel approach towards building Search Engines is using Evolutionary Algorithms was also discussed. Results were improved up to a certain extent by using “Evolutionary Algorithms” that try to achieve the best of both values by compromising the least.

In [21], to overcome the problem of increased complexity with justifiable effort in an acceptable time period, and to develop systems that solve these complex problems, new modified GA algorithms were required. Using these nature inspired search methods, it was possible to overcome some limitations of traditional optimization methods, and to increase the number of solvable problems.

In [22], since Ant Colony Algorithm might have produced redundant states in the graph, it’s better to minimize such graphs to enhance the behavior of the inducted system. Ants adaptively modify the way the problem was represented and perceived by other ants, but they were not adaptive themselves. The genetic programming paradigm permitted the evolution of computer programs which performed alternative computations conditioned on the outcome of intermediate calculations, which performed computations on variables of many different types, which can perform iterations and recursions to achieve the desired result, which defined and subsequently used computed values and sub-programs, and whose size, shape, and complexity were not specified in advance.

In [23], the performance of ABC algorithm with those of GA, DE, PSO and ES algorithms was compared on a large set of unconstrained test functions. From the results obtained in this work, the authors concluded that the performance of ABC algorithm was better than or similar to that of these algorithms although it used less control parameters and it efficiently used for solving multimodal and multidimensional optimization problems.

In [24], the reason of selecting these two search algorithms for empirical study lied in that, biological evolution of swarm intelligence; PSO is the effective algorithm in the field of swarm intelligence and ACO for its own advantages of stochastic search. PSO was more effective for the case of large candidate service number. An ACO with pheromone matrix and JPSO techniques have been adopted to solve the QoS multicasting routing problem in communication network. The solution generated by ACO is regulated by position update strategy of PSO, which extended search scope and increase avoids local optimization efficiently. The method of positioning update in JPSO was notified in order to adapt our discrete multi-objective multicasting routing problem. This proposed algorithm utilized PSO algorithm that has emerged as a new heuristic that can efficiently solve large-scale optimization problems. This study differed from existing literature in the following aspects: First, in this study various QoS measures were considered such as cost, bandwidth, delay and jitter. The proposed model treated these constraints separately, and extended to add more constraints. Second, new discrete PSO operators had been presented to modify the original PSO velocity and position update rules to the discrete solution space in the multicast routing problem. Third, a new adjustable PSO-ACO hybrid multicast routing algorithm which combines PSO with genetic operators was proposed. The performance of the adjustable hybrid model was optimized by two driving parameters that gave preference to either PSO or ACO. The proposed hybrid algorithm overcame the disadvantages of both PSO and ACO, and achieved better QoS performance.

V. CONCLUSION

In this paper, cloud computing technique is being reviewed. Also, load balancing in cloud computing have been discussed. The policies and some major goals of load balancing have also been discussed. Some of the meta-heuristic techniques are discussed for balancing the load in cloud computing.

ACKNOWLEDGMENT

This research was supported by MMU, Mullana. Firstly, I would like to thank Dr. Sandip Kumar Goyal, who is guiding me in my research work and moderated this paper and in that line improved the manuscript significantly. Also, I am grateful to Dr. V. K. Katiyar, who provided the useful resources, whenever required.

REFERENCES

Ant Colony Optimization
Fault Tolerance, R. D. Ramesh & D. Powar


Pooja Mangla is pursuing Ph.D from M. M. Engineering College, Maharishi Markandeswar University, Mullana, Ambala. She completed MTech from M. M. Engineering College, Maharishi Markandeswar University, Mullana, Ambala. She completed BTech from Kurukshetra University, Kurukshetra, Haryana. Her area of research is Load Balancing in Cloud Computing.

Dr. Sandip Kumar Goyal is HOD, CSE Department, M. M. Engineering College, Maharishi Markandeswar University, Mullana, Ambala. He received his B.Tech, M.Tech. from Kurukshetra University, Kurukshetra, India and is currently enrolled as a Ph.D. scholar in the department of Computer Science and Engineering at M.M.University, Mullana, Ambala, Haryana, India. He is presently serving as Assoc. Professor in Computer Engineering Department of M.M. Engineering College, Mullana, Ambala. He is in teaching since 2000. He has published more than 10 research papers in International and National journals and conferences. His research area Distributed Computing and Cloud Computing.