**Chapter 2 Literature Review**

*In this chapter we introduced our thesis literature, Cloud Computing, task scheduling and load balancing description. In section 2.1 we discussed about literature introduction; in section 2.2 we discussed thesis literature as Cloud Computing, Load Balancing; in section 2.3 we discussed about different types of algorithms and benefits of using these algorithms in cloud load balancing process; in section 2.4 we should summarize the chapter.*

### Introduction

A thesis is a statement in a non-fiction or a fiction work that a writer intends to support and prove. To understand our thesis work, literature review is important. Our work is all about compute the overall at auto scaling, task scheduling and finally get response time of the virtual machines. Our thesis work “Efficient Load Balancing In Cloud Environment With Modified Ant Colony Algorithm” named. To understand these first we have to understand these concepts. We discussed about these literatures.

### Cloud Computing

Our thesis named “Efficient Load Balancing In Cloud Environment With Modified Ant Colony Algorithm” is a work of developing a algorithm for balancing the load in cloud environment under the research area of Cloud Computing. In this section we should discuss about Cloud Computing, Load Balancing, Different types of Cloud Services, Load balancing methods.

#### Overview of Cloud Computing

Cloud computing is an emerging technology for providing effective and efficient computational services to many organizations worldwide. Services offered by the Cloud can be Software as a Service (SaaS), Platform as a Service (PaaS) or Infrastructure as a Service (IaaS) which can be offered in private, public or hybrid Clouds[1]. Cloud computing incorporates concepts of parallel and distributed computing to provide shared resources; hardware, software and information to computers or other devices on demand. These are emerging distributed systems which follows a “pay as you use” model. The customer need not buy the software or computation platforms. With internet facility, the customer can use the computation power or software resources by paying money only for the duration he/she has used the resource [2].

Many cloud providers are available nowadays like amazon web services, google cloud, etc. Cloud provider provides many services to the user which is Reliable, Efficient and Low cost.

Using virtualization technology, cloud data centers become more secure and flexible and provide better support for on-demand allocation.

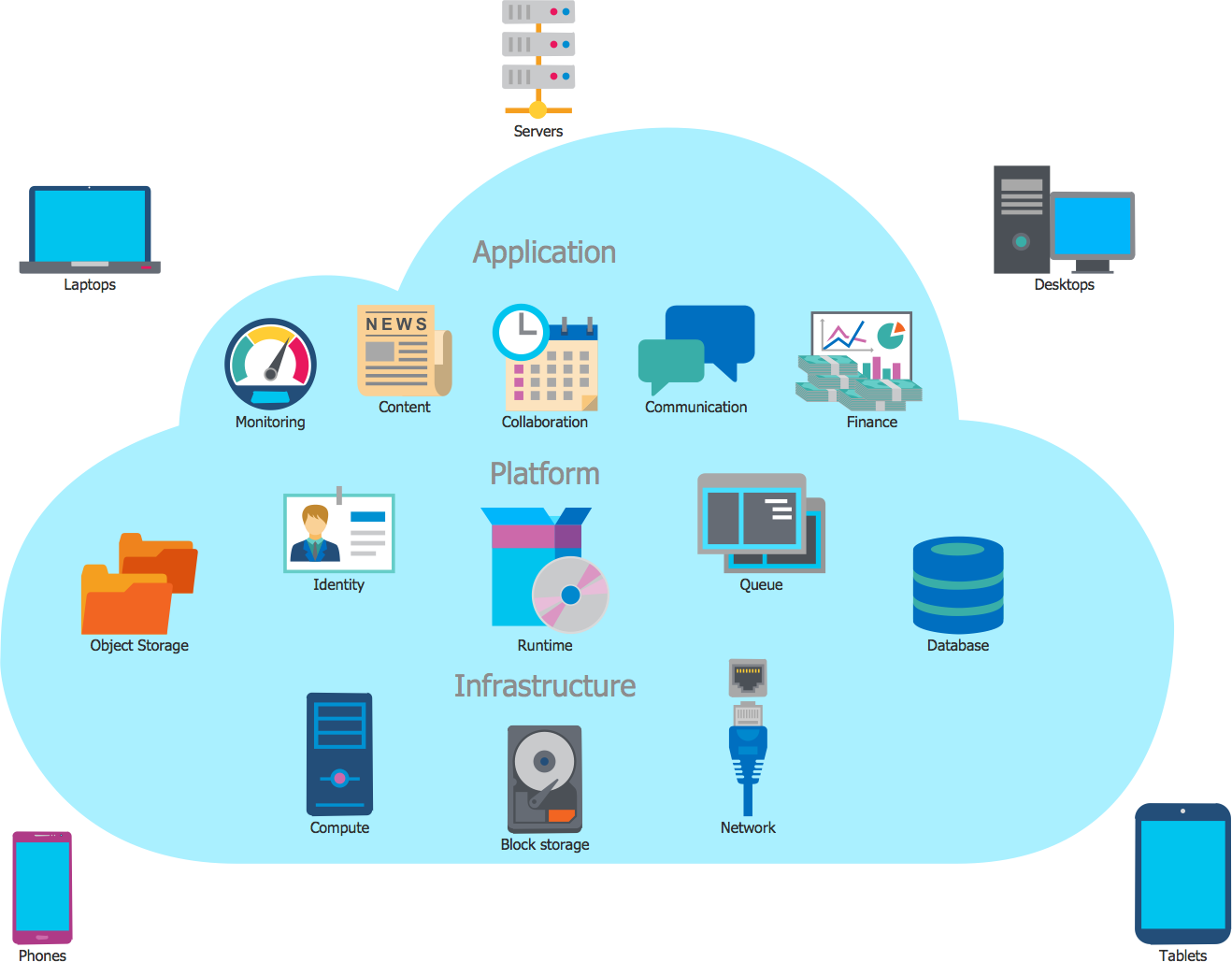


Figure 2. 1: Over View of Cloud Computing

###### Types of cloud computing deployment models

In cloud computing four types of deployment models are available-

1. Private Cloud: A private cloud is a cloud computing model which provides a cloud- based environment in which applications and resources can be access by the specified client only.
2. Public Cloud: It is a computing structure that is offered by the third party over the public internet which makes resources and application available for any user who wants to use them.
3. Community Cloud: A community cloud is an alliance effort which provides an infrastructure shared among many organizations that means the applications can be shared between them.
4. Hybrid Cloud: A hybrid clouds is a mix-up of two cloud that is a public and private cloud.

###### Types of Cloud service model

The Cloud provides three types of services models, these services modes are shown in the below figure:

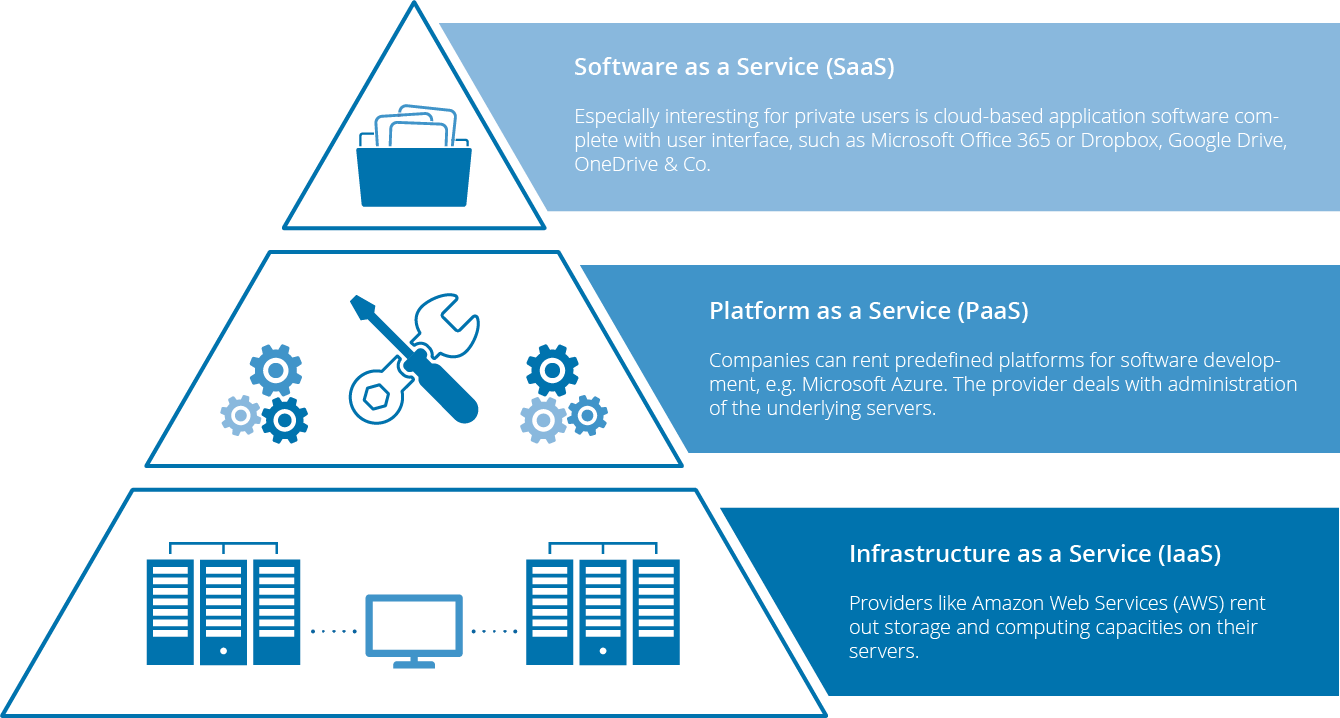


Figure 2. 2: Cloud Service Model

The models are described below- Software-as-a-Service

SaaS model provides the next features:

* access to the software from anywhere;
* payment for software usually include technical support;
* platform responsibilities managed by provider;
* sass applications are customizable and can be integrated with other business;
* anyone can pay monthly or only for the services you actually use;
* the provider performs maintenance and update of the software.

Platform-as-a-Service

PaaS model provides the next features:

* Simple. Cost-effective development and deployment of apps;
* anyone can use automated scalability of resources and do not need to allocate them manually;
* anyone can manage the services he/she develops and the provider manages all the rest.
* Significant reduction in the amount of coding needed

Infrastructure-as-a-Service

IaaS model provides the next features:

* anyone can rent all the tools (hardware and others) you need and use it how you need;
* anyone can receive administrator rights inside rented hardware;
* the MSP guarantees the availability of resources allocated to you;
* the provider manages all hardware issues.

###### Benefits of cloud computing

###### High Speed

Cloud computing allows you to deploy your service quickly in fewer clicks. This faster deployment allows you to get the resources required for your system within fewer minutes.

###### Back-up and restore data

Once the data is stored in a Cloud, it is easier to get the back-up and recovery of that, which is otherwise very time taking process on-premise.

###### Mobility

Employees who are working on the premises or at the remote locations can easily access all the cloud services. All they need is an Internet connectivity.

###### Unlimited storage capacity

The cloud offers almost limitless storage capacity. At any time you can quickly expand your storage capacity with very nominal monthly fees.

###### Collaboration

The cloud computing platform helps employees who are located in different geographies to collaborate in a highly convenient and secure manner.

**Cost Savings**

Cost saving is the biggest benefit of cloud computing. It helps you to save substantial capital cost as it does not need any physical hardware investments. Also, you do not need trained personnel to maintain the hardware. The buying and managing of equipment is done by the cloud service provider.

###### Automatic Software Integration

In the cloud, software integration is something that occurs automatically. Therefore, you don't need to take additional efforts to customize and integrate your applications as per your preferences.

###### Reliability

Reliability is one of the biggest pluses of cloud computing. You can always get instantly updated about the changes.

###### Strategic edge

Cloud computing offers a competitive edge over your competitors. It helps you to access the latest and applications any time without spending your time and money on installations.

###### Quick Deployment

Last but not least, cloud computing gives you the advantage of rapid deployment. So, when you decide to use the cloud, your entire system can be fully functional in very few minutes. Although, the amount of time taken depends on what kind of technologies are used in your business

###### Open research issues in Cloud Computing

Cloud computing is the fastest growing technology. That is why so many researchers are doing research on this filed. Open Research issues in cloud computing are listed below:

* + **Data Security**: Security is the greatest challenge or issue of cloud computing according to International Data Corporation (IDC). When we save our data or run our software into others hard disk using others CPU appears to be very risky. Organization’s data and software face serious risk of security issues like data loss, phishing, botnet etc.
  + **Load Balancing:** The second major open research issues in cloud computing is load balancing. In this issue many research is work for maintaining a load in virtual machine. Load balancing means distributed the upcoming request among the multiple servers to maintain the load.
  + **Disaster Recovery:** Disaster is a suddenly accident for short time but its caused large amount of damage in society. Disaster Recovery issues are data centers down during disaster, Data Backup, Cost, Failure detection and security**.**

**Availability of Service:** Availability refers to the ability of a user to access information or resources in a specified location and in the correct format. Simply availability mean services are available anytime and anywhere. The issues are protecting the confidentiality and integrity of data, Security and identify single point of failure.

### Cloud Load Balancing

In this section we discussed about Cloud Load Balancing, Types of Load Balancing process, Goals of Load Balancing and Different types of load balancing algorithms.

#### Overview of Load Balancing

Load Balancing is a computer networking method to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload [3]. Thus, load need to be distributed over the resources in cloud-based architecture, so that each resource does approximately the equal amount of task at any point of time.

Basic need is to provide some techniques to balance requests to provide the solution of the application faster. Cloud vendors are based on automatic load balancing services, which allow clients to increase the number of CPUs or memories for their resources to scale with increased demands. This service is optional and depends on the client’s business needs [4].

It has helped networks and resources, to provide maximum throughput with minimal response time. Cloud platforms enable enterprises to lease computing power in the form of virtual machines. Because hundreds of thousands of virtual machines (VMs) are used, it is difficult to manually assign tasks to computing resources in clouds [5]. So we need an efficient algorithm for load balancing in the cloud environment.

Load Balancing is one of the major issues related to cloud computing, the load may represent a CPU capacity, memory, network load etc. It is necessary to distribute the load equally among the nodes in a network. This results in agile and efficient performance of the system. Thereby it avoids heavily loading or under loading of nodes in a network [4].

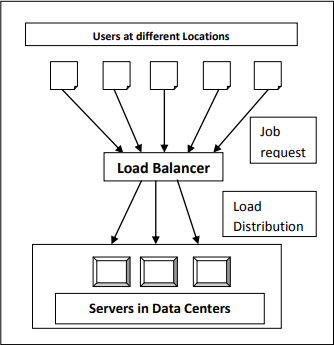


Figure 2.3. 2: Cloud Load Balancer

#### Metrics for Load Balancing in Cloud

Various metrics considered in existing load balancing techniques in cloud computing are discussed below [4] [3]-

* + **Scalability** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.
  + **Resource Utilization** is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.
  + **Performance** 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.
  + **Response Time** is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.
  + **Overhead** Associated 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.

#### Load Balancing on the basis of Cloud Environment

Cloud computing can have either static or dynamic environment based upon how developer configures the cloud demanded by the cloud provider [5].

**Static Environment:** In static environment the cloud provider installs homogeneous resources. Also the resources in the cloud are not flexible when environment is made static. In this scenario, the cloud requires prior knowledge of nodes capacity, processing power , memory, performance and statistics of user requirements. Round Robin algorithm provides load balancing in static environment [5].

**Dynamic Environment:** In dynamic environment the cloud provider installs heterogeneous resources. The resources are flexible in dynamic environment. In this scenario cloud cannot rely on the prior knowledge whereas it takes into account run-time statistics. The requirements of the users are granted flexibility (i.e. they may change at run time). Algorithm proposed to achieve load balancing in dynamic environment can easily adapt to run time changes in load. Dynamic environment is difficult to be simulated but is highly adaptable with cloud computing environment**.** LBMM (Load Balancing Min-Min) algorithm uses for resource allocation in dynamic environment [5].

#### Goals of Load Balancing Algorithms

The goals of load balancing are to [6] [7]:

* + For better performance.
  + Manages resources efficiently.
  + For achieving Stable System state.
  + For accompanying further modifications.
  + For construction of fault tolerant system
  + To have the ability to adjust itself in accordance with any modifications
  + Utilizes each the systems resources as efficiently as possible.

#### Different types of Load Balancing Algorithms

* + **Round Robin Algorithm:** In this algorithm [4] [5], the processes are divided between all processors. Each process is assigned to the processor in a round robin order. The process allocation order is maintained locally independent of the allocations from

remote processors. Though the work load distributions between processors are equal but the job processing time for different processes are not same. So at any point of time some nodes may be heavily loaded and others remain idle. This algorithm is mostly used in web servers where Http requests are of similar nature and distributed equally.

* **Equally Spread Current Execution Algorithm**: Equally spread current execution algorithm [8] process handle with priorities. it distribute the load randomly by checking the size and transfer the load to that virtual machine which is lightly loaded or handle that task easy and take less time , and give maximize throughput. It is spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines.
* **Throttled Load Balancing Algorithm**: Throttled algorithm [9] is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access that load easily and perform the operations which is give by the client or user. In this algorithm the client first requests the load balancer to find a suitable Virtual Machine to perform the required operation.
* **BEE FORAGE BEHAVIOUR:** The artificial bee colony formula (ABC) algorithm supported the intelligent forage behavior of honey bee swarm and was planned by Karaboga in 2005. The formula is totally galvanized by natural forage behavior of honey bees [10].
* **ANT COLONY OPTIMIZATION:**nt algorithm’s is a multiagent approach to difficult combinational optimization problems. Example of this approach is travelling salesman problem (TSP) and the quadratic assignment problem (QAP). These algorithms were inspired by the observation of real ant colonies. Ant’s behavior is directed more to the survival of the colonies [11]. They not think for individual.

#### Challenges in Cloud Computing Load Balancing

Here we discuss the challenges to be addressed when attempting to propose an optimal solution to the issue of load balancing in Cloud Computing [9]:

**Spatial Distribution of the Cloud Nodes:** Some algorithms are designed to be efficient only for an intranet or closely located nodes where communication delays are negligible However, it is a challenge to design a load balancing algorithm that can work for spatially distributed nodes.

**Storage/ Replication:** A full replication algorithm does not take efficient storage utilization into account. This is because the same data will be stored in all replication nodes. Full replication algorithms impose higher costs since more storage is needed.

**Algorithm Complexity:** Load balancing algorithms are preferred to be less complex in terms of implementation and operations. The higher implementation complexity would lead to a more complex process which could cause some negative performance issues.

**Point of Failure:** Controlling the load balancing and collecting data about the different nodes must be designed in a way that avoids having a single point of failure in the algorithm. Some algorithms (centralized algorithms) can provide efficient and effective mechanisms for solving the load balancing in a certain pattern. However, they have the issue of one controller for the whole system. In such cases, if the controller fails, then the whole system would fail. Any Load balancing algorithm must be designed in order to overcome this challenge.

.

### Related Works

Load balancing mechanism distributes the work across multiple computing resources to utilize them effectively and at the same time eliminating a condition during which bound nodes are over loaded whereas others are beneath loaded.

Tasks that have the minimum expected time for completion are deduced and assigned to the corresponding machine.

Kokilavani et al. [12] proposed a MinMin (LBMM) algorithm primarily for load balancing which enhances resource utilization and reduces makespan. This has been done in two phases: the former phase executes the conventional Min-Min algorithm, while the later reschedules of tasks for effective resource utilization The work proposed by Lorpunmanee et al. [13] proposed the scheduling problem by introducing a model based on dynamic information on grid using ACO algorithm.

Load balancing under tri-level cloud computing was introduced in [14] by combining Load Balancing Min-Min (LBMM) and Opportunistic-Load Balancing (OLB) to maintain the load balancing of system and enhance efficiency for execution. Nowadays, it is important to consider multiple objectives simultaneously

Load balancing is a method that has helped networks and resources, to provide maximum throughput with minimal response time. Cloud platforms enable enterprises to lease computing power in the form of virtual machines. Because hundreds of thousands of virtual machines (VMs) are used, it is difficult to manually assign tasks to computing resources in clouds [15]. So we need an efficient algorithm for task scheduling in the cloud environment.

.