1.1. Preamble

Cloud computing has become a standout amongst the most fascinating specialized fields in the modern era. It has shown its effect on data stockpiling, information technology, programming designing, and organizations. The National Institute of Standards and Technology (NIST) characterizes cloud computing as: “the cloud computing is a paradigm to enable access for resources pooling, convenient, on-demand, ubiquitous, which can be easily delivered with different types of service provider interaction” [1].

Users can access the resources of the cloud from any place in the world according to the request. The Cloud Service Providers (CSPs) have a huge number of data centers which are incentivized by the profits from users who pay to get service access, whereas the users are attracted through the opportunity of minimizing the charges for complete implementation of those services [2]. Cloud data centers consist of a large number of servers. A large number of users share the resources in data center through virtualization [3], which make cloud data centers more flexible, so as to offer better support for on-demand provision. Virtualization hides heterogeneity of server, enhanced server utilization, and enables consolidation of the server. Each server hosts a large number of heterogeneous Virtual Machines (VMs) and receives variable and unpredictable load which may cause an imbalance in resource usage [4].

In this system, the Quality of Service (QoS) requirements is guaranteed to the user through established contracts between the service provider and the user according
to Service Level Agreements (SLAs). Pay-as-use and QoS concepts are used in cloud computing to provide utility grid characteristics over the Internet, meaning the cloud computing is an extension of the utility grid. Pay-as-use and on-demand usage of cloud resources make this environment cost-effective and more scalable. Moreover, users can rent this resource as required instead of being concerned of their maintenance and so on [5]. Therefore, providing cloud services that fulfill users’ requirements without exceeding SLA is a major challenge. Currently, services of the cloud are provided and scheduled depending on the resource availability, but the anticipated performance is not assured. Hence, service providers should improve their ecosystems to guarantee the QoS requirements of all the cloud components [6].

The data center which is the main entity in the cloud receives a large number of users’ requests that come from anywhere in the world, and each task should be processed. Therefore, each cloud provider requires a scheduling strategy in order to schedule and execute tasks. Several cloud providers may have their own scheduling policies regarding the scheduling problem of assigning the tasks to appropriate VMs, which are created over the clouds [7].

As with all task scheduling problems, optimal task scheduling in the cloud is classified as a NP-complete problem. In this payment model, one of the major essential components which are mulled over in tasks scheduling methodologies is execution cost, in addition to execution time. In this way, compatibility between two of the most noteworthy variables of QoS, execution time and execution cost makes the scheduling tasks issue on the cloud significantly more complex and difficult [5].
1.2. Research Problem

The number of cloud users increased dramatically day by day with the flexibility offered by service providers. The providers focused on improving the capacity of their resources like infrastructure for accommodating large range of users. But, there is no service provider having unlimited resources to address peak or fluctuating user requests. Also, the service providers need to assure that there is adequate number of resources to achieve the requirements of QoS such as execution time, deadline constraints, and budget constraints that are ensured to users.

Anyway, implementing too many tasks or demands on a single resource will create tasks interference with each other. So, the performance will become low and unpredictable, causing discouragement to the users. The complex issue is concluding in assigning and executing the workloads to suitable resources in the cloud environment.

The cloud service providers earn revenues when the user’s demand is completed successfully. On the other hand, a provider pays a penalty cost on SLA violation.

Each service provider owns especial SLA which may vary from those of other service providers. However, lesser number of resources should be used in executing the tasks to preserve an acceptable QoS level or reduce the completion time of the tasks. These issues are solved by providing the best task scheduling mechanism which determines the effective resources along with considering the significance of QoS metrics as the execution cost with budget and execution time with a deadline, in other words, achieving the satisfaction of the user and service provider.
1.3. Research Questions

1. How to build a scheduling model to utilize resources efficiently during load management and reduce the total completion time while fulfilling deadline constraint?

2. How to formulate a model which reduces the total cost for task scheduling problem in a cloud computing environment, which consists of the cost of processing, memory, bandwidth, and storage based on budget constraint?

3. How can we design task scheduling model to execute tasks while achieving user defined constraints (deadline and budget)?

1.4. Research Objectives

1. To build a scheduling model capable of utilizing the resources efficiently during load management by reducing the total completion time for the longest task (makespan) based on deadline constraint.

2. To formulate a model to reduce the total cost for task scheduling problem in a cloud computing environment, which consists of the cost of storage, cost of processing and cost of communication based on the budget constraint.

3. To design a model capable of executing tasks based on both deadline constraint and budget constraint. The performance of the proposed models is evaluated by comparing with other applicable algorithms.
1.5. **Scope of Research**

1. Define the task scheduling problem in cloud computing environment for exploiting resource utilization and get high-performance from cloud computing system.

2. Address the gap in the task scheduling problem by reducing the execution time of tasks and cost of used resources in cloud computing based on some constraints.

3. Use SLA to determine QoS metrics such as total completion time (makespan), response time, etc., to measure the service offered by the service provider to the users.

4. Include a comparative study of some scheduling strategies in cloud computing and demonstrate the results of our proposed models being better, compared to other algorithms.

1.6. **Significance of our Research**

Cloud computing is a paradigm of distributed and parallel systems. It involves collection of heterogeneous interconnected datacenters which depend on virtualization technique and provisioned as a dynamically to the user through a negotiation between the cloud providers and cloud users. With increasing cloud users and the limited resources (e.g., computing capacity, bandwidth) in the cloud, guaranteeing the QoS and improving resource utilization are still essential challenges in cloud computing. Often, the users specify constraints such as deadline and budget for scheduling their tasks over the resources in the cloud, where each service provider has several resources that have different cost. However, the fast resources are expensive as compared to cheap resources. A trade-off solution is needed in these complex conflicts between constraints.
and dynamic nature of the cloud. The main issue of the significance of this research is how to get efficient task scheduling in cloud computing environment and achieve a balance between time and cost.

This research manages the models which execute tasks in proposed scheduling strategies in view of different constraints to satisfy the user satisfaction besides the service provider satisfaction.

A scheduling strategy will be presented to reduce the total completion time (makespan) and response time in the cloud computing environment to get high-performance during task execution in heterogeneous resources.

1.7. Motivation of Research

The motivation for using cloud computing is to use the services (software as services, platform as services, and infrastructure as services) without the need to establish an information technology base to provide those services. Payment will be made for what has been utilized from these resources.

The important requirement in cloud computing environment is to meet QoS requirements. The quality of service is defined based on the service level agreement. The agreement between the user and service provider will determine the quality of services such as tasks execution time, and cost of utilized resources.
1.8. **Research Contributions**

This research contributes the best solutions for solving the task scheduling problem and face its challenges. Our contributions are as the following:

1. We propose a model based on the deadline constraint, capable of assigning the tasks while adapting to the user and provider satisfaction in the following manner:
   
   i. The tasks are ranked according to the length of tasks priority in ascending order.
   
   ii. The VM’s state is labeled as successful when it achieves the deadline constraint.
   
   iii. Selecting a suitable VM which returns minimum processing time of the task. The details of this model will be discussed in Chapter 3.

2. A budget distribution strategy that allocates the tasks according to their characteristics in order to facilitate resource selection decisions. This strategy is implemented as follows:
   
   i. At first, the VMs which achieve task priority and budget are determined.
   
   ii. The task attributes (length, file size) are detected.
   
   iii. The task is assigned to the VM which fulfills the user-defined constraint. This model is explained in Chapter 4.
3. We have proposed a novel model, where the user defines two constraints - deadline and budget as follows:

i. Users’ tasks will be assigned to appropriate VM based on user-defined constraints.

ii. Achieving user satisfaction through minimizing the completion time and cost of this process by implementing the tasks onto VMs and meeting the QoS constraints.

iii. Provider satisfaction is achieved through maximizing profit revenue and resource utilization. The explanation of this model is provided in Chapter 5.

1.9. Research Methodology

In this research, tasks are assigned and executed on the available resources provided within specific constraints. The constraints used in this research are deadline and budget.

Several algorithms are proposed to handle these types of issues, but most of them do not cover the metrics that are needed to evaluate the performance of task scheduling. Therefore, we propose to achieve effective performance in task scheduling mechanism with our models.
1.9.1. The Proposed Models’ Framework

The proposed models in our research are as follows:

i. The first model is the Deadline-Aware Priority Scheduling (DAPS) model. In this model, the tasks are scheduled and assigned to the appropriate resources so as to reduce the makespan under the deadline constraint and to reduce minimum completion time. One of the criteria that must be taken into consideration when designing an effective task scheduling model is user satisfaction. The methodology relates to focus on scheduling the tasks based on the deadline constraint, and so, deadline is assumed as the execution time of the task.

ii. The second model is the Budget-Aware Scheduling (BAS) model, where scheduling the tasks and its assignment is done on the available resources with reducing the makespan under the budget constraint and the cost of resources that are used is calculated. The service provider offers services and resources to the user, so the user decides the resources he needs and then pays for the costs of what was consumed based on the budget constraint.

iii. Finally, the Deadline Budget Scheduling (DBS) model. In this model, scheduling tasks and assigning them to the heterogeneous available resources is done with two conflicting QoS requirements: time and cost while meeting user satisfaction. The most significant factors of the proposed DBS model are to minimize the makespan under user-defined deadline and reduce monetary costs while not surpassing the user-defined budget.
Through the proposed models, we were able to meet the user's desire and obtained satisfaction of the provider and the user by providing different models for scheduling and executing the tasks based on different constraints. Figure 1.1 shows the sequence of the proposed models.
### Figure 1.1: Proposed models of our research

<table>
<thead>
<tr>
<th>Deadline-Aware Priority Scheduling (DAPS) model</th>
<th>Budget-Aware Scheduling (BAS) model</th>
<th>Deadline Budget Scheduling (DBS) model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tasks with Deadline</strong></td>
<td><strong>Tasks with Budget</strong></td>
<td><strong>Tasks with Deadline &amp; Budget</strong></td>
</tr>
<tr>
<td>VM State</td>
<td>VM Checking</td>
<td>Scheduler Detector</td>
</tr>
<tr>
<td>Estimated Completion Time</td>
<td>Checking Task Attributes</td>
<td>Determination</td>
</tr>
<tr>
<td>Minimum Completion Time</td>
<td>Calculate Resources Cost</td>
<td>Calculate Resources Cost</td>
</tr>
<tr>
<td>Mapping to VM and VM Processing Time</td>
<td>Task mapped to VM to achieve the task’s requirements</td>
<td>Calculate Completion Time, Communication Time, and Cost of Resources</td>
</tr>
</tbody>
</table>

**Cloud Users**

**Cloud Service Provider**

**Available Resources**
1.9.2. Task Scheduling Steps

In cloud computing, the user sends the request or task, which may contain different information about the user's requirements such as a constraint, priority, and so on. On the other side at the service provider that owns a special task scheduling, the scheduler receives the requests from the users to schedule their tasks as agreed upon in the service level agreement contract between the users and the cloud service provider, in order to guarantee the quality of service while getting a revenue profit from the services used.

The scheduler selects the suitable resources from the heterogeneous cloud resources to execute these tasks based on some constraints, and the schedule here is represented by the proposed model and is the intermediary between the users and the service providers.

The steps of task scheduling are presented in Figure 1.2.
1.9.3. **Simulation Tool**

To simulate the task scheduling in homogeneous and heterogeneous cloud computing environment, there is a need for a tool which supports this environment in terms of flexibility and receives increased number of user requests with increased use of resources.

CloudSim toolkit is one of the best open source toolkits that supports the creation of data center and Physical Machines (PMs), as well as the virtual machines.
The CloudSim toolkit is a simulator as an event driven and is hosted in Java language. CloudSim toolkit allows to define and extend the policies in all the CloudSim components. So, it is considered as a good tool for research, which can simulate the complexities emerging from the environments [4].

The simulation experiments were run on a laptop with the following configurations: 2.5 GHz, Corei5 CPU, 4 GB memory, and 512 GB hard disk.

1.9.4. Datasets

In the cloud computing environment, the datasets used are different according to the applications used. In this research, two types of dataset are used, the first type is taken from the NASA site for research and the second type is taken by using the CloudSim toolkit through which the dataset is generated. The details of each of these types of dataset used will be explained in chapters 3, 4, and 5.
1.10. Thesis Organization

This thesis is organized into six chapters as mention in Figure 1.3. Chapter one offers an overview of the prologue. The remaining chapters are presented as:

Chapter 2: This chapter, is divided into two sections. Section 1 is about the basics of cloud computing namely: clarification of cloud computing properties, kinds of cloud service models, deployment models, virtualization in a cloud system, clarification of the Service Level Agreement between cloud users and cloud providers, and also Quality of Service are discussed. Finally, the challenges of cloud computing are presented. Section 2 clarifies the task scheduling in cloud computing, task scheduling algorithms, task scheduling level, task scheduling based on QoS, task scheduling metrics, recent researches on task scheduling, and task scheduling challenges.

Chapter 3: In this chapter, the proposed deadline-aware priority scheduling model is discussed, including deadline-aware priority scheduling model methodologies and implementation. The evaluation of this model demonstrated the efficiency of our proposal.

Chapter 4: Here, the proposed budget-aware scheduling model is illustrated extensively, the ecosystem of the model and its effectiveness on the user’s budget when executing the tasks on cloud resources based on budget constraint. The implementation and evaluation of this model proved the efficiency of our work.

Chapter 5: A proposed model named deadline budget scheduling algorithm is presented which allows cloud users to take into account two constraints, deadline and budget, when executing their tasks over cloud resources. Finally, we have tested and evaluated our proposal to prove our system efficiency.
Chapter 6: This chapter summarizes the entire thesis work, and the scope for future work.

1.11. Chapter Summary

This chapter gives a brief overview of the research work. The research problem is reviewed in detail and discusses the need to find solutions that contribute to solving this problem effectively.

The research questions and objectives of this research are formulated. Through achieving the objectives, we got the answers to the research questions. The significance and scope of the research to reach and overcome most of the problems in task scheduling algorithms. Then, the motivation of the research is clearly identified. Finally, the researcher's contributions clarify that the methodologies used to solve the research problem are better compared to other algorithms which are clear proofs of our system performance.

Figure 1.3 summarizes the organization of this thesis.
Figure 1.3: Thesis organization